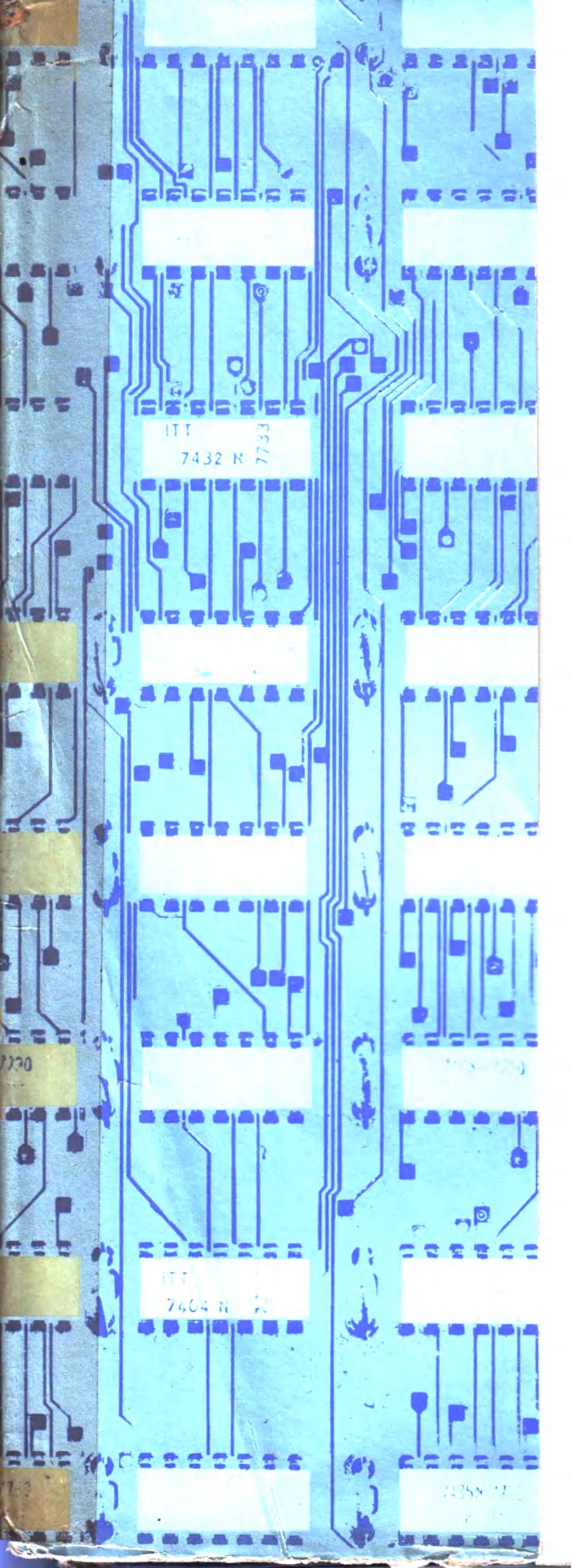
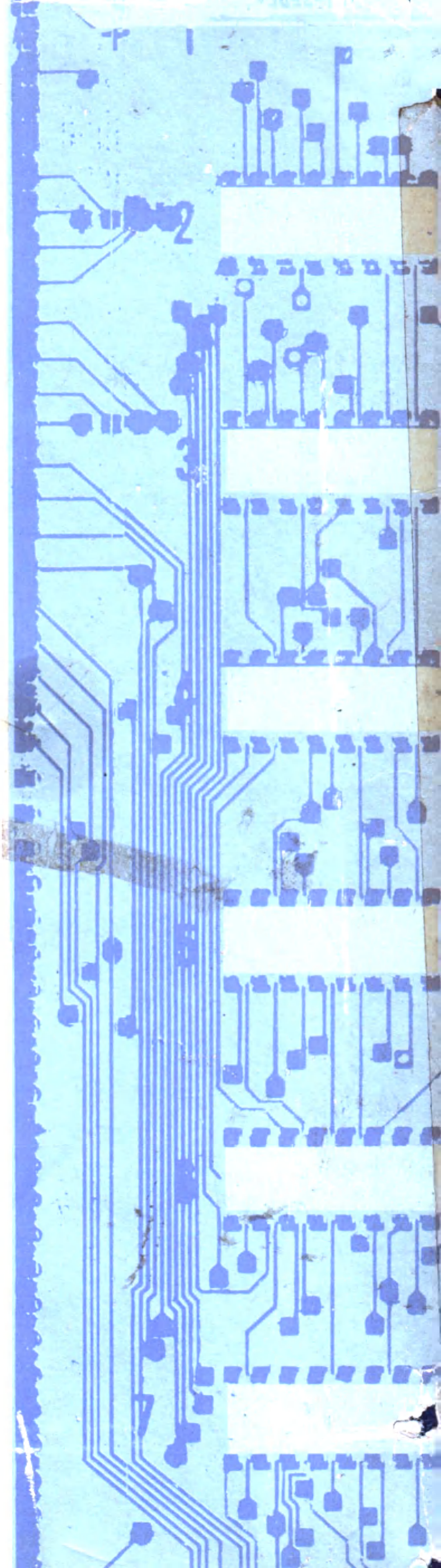
I. Zherebtsov

basic electronics

Mir Publishers
Moscow


ITT
7432 N
7733
ITT
7404 N

Mir Publishers of Moscow publishes Soviet scientific and technical literature in twenty four languages including all those most widely used. Mir translates texts into Russian, and from Russian originals produces books in English, German, French, Italian, Spanish, Portuguese, Czech, Slovak, Finnish, Hungarian, Mongolian, Arabic, Persian, Hindi, Tamil, Kannada, Vietnamese, Dari, Laotian, Khmer, Bengali, Marathi, and Telugu. Titles include textbooks for higher technical and vocational schools, literature on the natural sciences and medicine (including textbooks for medical schools), popular science and science fiction. The contributors to Mir Publishers' list are leading Soviet scientists and engineers from all fields of science and technology, among them more than forty Members and Corresponding Members of the USSR Academy of Sciences. Skilled translators provide a high standard of translation from the original Russian. Many of the titles already issued by Mir Publishers have been adopted as textbooks and manuals at educational establishments in France, Switzerland, Cuba, Syria, India, Brasil and many other countries. Mir Publishers' books in foreign languages can be purchased or ordered through booksellers in your country dealing with V/O "Mezhdunarodnaya Kniga", the authorised exporters.



The subject-matter of the book covers a wide range of material specific to electronics—from the basic principles underlying it to sophisticated devices employed in a multitude of applications. Among other things, there is a fairly detailed discussion of semiconductor materials and devices, electron tubes, photocells, optoelectronic devices, and integrated circuits. The text is liberally illustrated and includes a discussion of reliability and testing.

The book has been conceived as an aid in the study of electronics by college students, those relying on self-education, and hobbyists.

Docent Ivan P. Zherebtsov, Cand. (Pedagog. Sc.) taught electrical engineering and radio engineering at secondary educational establishments since 1928, has been lecturing on these subjects at colleges since 1946. Is a leading Soviet authority in the field of telecommunications. Has penned over 40 books and booklets many of which have been translated into foreign languages or published outside the Soviet Union. The most important of them are *Radio Engineering* (five editions), *Basic Electronics* (four editions), *Electric and Magnetic Circuits* (two editions), *An Introduction to UHF-SHT Radio Engineering* (three editions). Honorary member of the A. S. Popov Scientific and Technical Society on Radio, Electronics, and Communications.

I. Zherebtsov

basic electronics

Basic Electronics

И. П. Жеребцов

Основы электроники

Энергоатомиздат
Ленинград

I. Zherebtsov

Basic Electronics

ИЗДАТЕЛЬСТВО МИР

Mir Publishers Moscow

First published 1988
Revised from the 1985 Russian edition

Translated from the Russian by Boris V. KUZNETSOV

На английском языке

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-000021-6

Contents

Preface

Advances in science and technology today are inseparably linked to major breakthroughs in electronics, among them the advent of basically new designs of electron devices, both tubes and semiconductor components. A person who wishes to be well versed in state-of-the-art electronics, must above all learn how such devices operate, what characteristics and parameters they have, and how they can be used in electronic equipment. All of these matters are taken up in this book which has been written as an aid for college students, those studying telecommunications on their own, and radio amateurs.

At this writing, the number of electron device types is too large to be covered in a single book. An attempt to do would have resulted in either an unwieldy collection of data or an inadequate or even very scanty description of many devices. Quite naturally, the author has left out devices used on a very limited scale, if available commercially at all.

Semiconductor devices are discussed first because they are the heart of the equipments and circuits that most of those concerned with electronics have to do with, and for them a discourse on electron tubes would have been unnecessary.

Special emphasis in the book is placed on the matters that the author believes to be especially important. Apart from a description of the devices as such, he touches upon some of their uses. Otherwise, the reader would have acquired an incomplete idea about the component side of electronics.

This is an English translation of the fourth Russian edition. As compared with its predecessors, it has been extended to include chapters on microelectronics, photoelectronics, and optoelectronics. More material has been added on semiconductor devices while the presentation of tubes has been curtailed. Only a very brief mention is made of the devices that have fallen out of use in electronics, such as gas-filled rectifier diodes, hot-cathode thyratrons, and some others.

I wish to express my gratitude to Docent G. A. Fedotov, Cand. Sc. (Tech.), for his very careful review of the manuscript and his valuable suggestions.

Undoubtedly, the criticism voiced by N. V. Parol and V. A. Terekhov on the previous editions has served to improve the encouragement that has come from other people, and to all these people I want to express a heartfelt “Thank you”.

I. Zherebtsov

Introduction

I-1 Electronics Defined

Electronics is a field of science and technology concerned with the study, design, and use of devices that depend on the conduction of electricity through a vacuum, gas, or semiconductor.

For purposes of discussion it appears convenient to class it into *physical electronics* – the division that has to do with electron and ion processes occurring in a vacuum, gases, and semiconductors, and also at the interface between a vacuum or a gas and a solid or liquid, and into *electronic engineering* which has to do with the design and use of the devices that depend for their operation on the above processes. There is a special field called *industrial electronics*. As its name implies, it has to do with the design and use of electron devices in industrial applications.

Advances in electronics have largely been prompted by those of telecommunications, notably radio engineering. For this reason, it was at some time customary to call the two fields collectively as radioelectronics. Electron devices are the basis of telecommunication equipment and have a direct bearing on its performance. On the other hand, many of the problems that first arose in telecommunications later led to the advent of new and an improvement of existing electron devices. These devices are used in radio communication, television, sound recording and reproduction, radar, radio navigation, radiotelecontrol, and elsewhere. To this we should add the fact that electronics has penetrated other divisions of today's science, technology, and industry. Electron devices are doing many jobs in automatic control, teleoperation, wire (or line) communication, cinematography, nuclear engineering, rocketry, astronomy, meteorology, geophysics, medicine, biology, physics, chemistry, metallurgy, mechanical engineering, measurement and instrumentation, etc.

Progress in electronics has been a big help to cybernetics – the science and technology concerned with the study of control and information flows in artificial and natural systems, and has also served as a basis for high-speed electronic computers. Without electronics, man would not have been able to explore outer space with his probes, artificial Earth satellites, space vehicles, and unmanned interplanetary stations.

Electron devices provide powerful tools for studies and measurements, notably those which have, as such, nothing to do with electronics. Among them are electronic amplifiers, oscillators, rectifiers, oscilloscopes, and measuring instruments. Apart from research and automatic control, they come in useful in running a great variety of production processes. Electronics-based techniques have advanced our knowledge about the properties of many substances existing in nature, provided a deeper insight into the structure of matter, and have brought us closer to a proper understanding of the laws that govern the material world.

I-2 A Brief Historical Outline

The foundations for electronics as we know it today were laid by physicists back in the 18th and 19th centuries. A big impetus was given by the electronic theory of metallic conduction developed by many outstanding scientists in the late 19th and the early 20th century.

In 1887, Heinrich Hertz of Germany, known for his experiments with electromagnetic waves, discovered the photoelectric effect. In 1888, Alexander Stoletov of Russia investigated Hertz's discovery and formulated the laws of the photoelectric effect, thus breaking ground for the design and use of photoelectronic devices. Also in 1888. Vladimir Ulyanin of Russia built the first selenium photocells. It should be noted that it was not until 1904 that the photoelectric effect was explained by Albert Einstein who worked on a theory that the radiant energy could only be transferred in discrete packets called photons.

The year 1883 saw the discovery of thermionic emission by Thomas A. Edison of the United States. In fact, it was first called the Edison effect. Unfortunately, Edison knew nothing about electrons and could not explain what he had observed. For the first time, an in-depth study of thermionic emission was made by O. W. Richardson of Britain, who was the first to derive, in 1912, a thermionic electron emission equation based on the classical electronic theory of metallic conduction. (In 1923, S. Dushman applied the quantum theory to this same problem and derived his own version of the thermionic electron emission equation.) In 1897, Karl Braun of Germany built the first cold-cathode ray tube.

The use of electron devices for radio communication began in 1904 when Sir John A. Flemming of Britain produced a vacuum hot-cathode diode tube to rectify (detect) electromagnetic waves in radio receivers. He named his device a thermionic valve (for the reason that it permits only a unilateral flow of particles from the negative to the positive electrode, much as a mechanical valve does so to a flow of liquid or gas). For this device he obtained a patent in 1904 – this was the first electron tube.

At about the same time, A. Wehnelt of Germany discovered and investigated the increased electron emission by wires given a coat of alkali-earth metal oxides. His discovery ultimately led to the use of what has come to be known as the oxide cathode widely employed in state-of-the-art electron tubes. In 1905, A. Hull of the United States invented the gas-filled rectifier diode, another important milestone in the progress of electronics.

In 1906, De Forest of the United States made a discovery that is rated by many as one of the greatest engineering breakthroughs of modern times, the one that gave rise to the field of electronics. He observed that current flow in a diode could be controlled by the field produced by a grid of fine wires placed as a third electrode between the cathode and the anode (or plate). De Forest's original "Audion", as he called his invention, was the forerunner of the modern vacuum triode. In 1907, B. Rosing of Russia proposed to use a cathode-ray tube for image reception and later proved the viability of his invention. This places him among the originators of present-day television.

In 1909-1911, V. Kovalenkov of Russia built triodes adapted to service in long-distance telephone repeaters. A bit later, he added a second grid to the tube to produce a tetrode, that is, a four-electrode tube, which he likewise used in long-distance telephone repeaters. Similar four-electrode, or double-grid, tubes were built by Irving Langmuir of the United States somewhat later.

In 1913, Alexander Meissner of Germany (born in Vienna) was the first to use a vacuum triode as a self-excited vacuum-tube signal generator involving feedback. This had a decisive influence on the progress of radio engineering, especially in Europe; until then, undamped high-frequency oscillations had been obtainable only by means of alternators coupled to frequency multipliers or by utilizing the arc discharge as a type of negative resistance.

In Russia, the first triodes for the reception of radio signals were independently built by N. D. Papalexi and M. A. Bonch-Bruyevich in 1914-1916. In 1918, Bonch-Bruyevich headed a team at the Nizhny Novgorod Radio Laboratory in Russia to develop high-power transmitting and low-power receiving tubes. Valuable contributions were made by B. A. Ostroumov, A. M. Kugushev, N. A. Nikitin, and P. A. Ostryakov among many others.

In 1918-1919, Bonch-Bruyevich published his triode theory which played an important role in the design of vacuum tubes during the subsequent years. He also advanced a theory that explained signal amplification by the vacuum triode. Similar works better known to the Western world were independently published by Heinrich Barkhausen. In 1911, he took over the chair of communications engineering in Dresden where he founded the first institute on this subject in Germany. He made important contributions to the theory of nonlinear switching elements, formulated the electron-tube coefficients (and the equations relating them) that are still in use, and wrote a four-volume text on electron tubes. About the same time, W. Schottky of Germany added what has come to be known as the screen grid to the tube, thus producing the screen-grid tetrode. Today it is mainly of historical interest, but it was an important step in the development of the pentode, the most extensively used of all types of vacuum tubes.

Special mention should be made of the water-cooled high-power transmitting tube invented by Bonch-Bruyevich and developed much later outside the Soviet Union. Important steps were demountable transmitting tubes devised by A. L. Mints, N. N. Oganov and A. M. Kugushev also at Nizhny Novgorod. A team under V. P. Vologdin came up with several designs of high-power mercury-vapour rectifier tubes.

Large-scale R & D work in the field of electronics went on in Leningrad. Among the leading figures there was A. A. Chernyshev who invented the indirectly heated cathode in 1921.

In 1922, O. V. Losev at Nizhny Novgorod discovered that oscillations could be generated and amplified by a crystal (semiconductor) detector. He also observed the glow discharge at the detector's contact. Unfortunately, his findings were not followed up, and the inventor himself died during the Leningrad Siege. For a long time, work in this field was limited to theoretical studies of semiconductors and the design of semiconductor rectifiers.

Beween 1920 and 1930, much headway in the field of electron devices was made outside the Soviet Union. In 1926, A. Hull of the United States made far-reaching improvements in the screen-grid tetrode and in 1930 he took out a patent on the pentode which is, as already noted, the most extensively used of all types of vacuum tube. Improvements were made in gas-filled rectifier diodes, and the thyratron (a gas-filled triode) was invented. The next decade saw an impressive number of important discoveries and inventions in the field of electronics. In 1930, L. A. Kubetsky of the Soviet Union invented the photomultiplier (also known as the electron multiplier) later radically improved and commercialized by S. A. Vekshinsky and P. V. Timofeyev of the Soviet Union. In the United States, similar devices were produced by Farnsworth. In 1930-1931, A. P. Konstantinov and S. I. Katayev of the Soviet Union, working independently of each other, came up with the idea of television pick-up (camera) tubes. In the United States, Vladimir K. Zworykin engaged in investigations in the field of photoelectric emission and television. These studies led to his conception of a new type of television pick-up tube, the iconoscope, which he developed into a form suitable for practical picture transmission. Zworykin's second major step towards all-electronic television was the development of the kinescope or the television picture tube. All this had opened up broad vistas for rapid advances in the practical use of television.

In 1933, P. V. Shmakov and P. V. Timofeyev of the Soviet Union proposed the image iconoscope (or the superemitron), a far more sensitive TV camera pick-up tube with which the scene to be televized needs no strong lighting. In 1939, G. V. Braude of the Soviet Union proposed a still more sensitive TV camera pick-up tube later called the image orthicon. Also in the 1930s, experiments were made with very simple TV camera tubes known today as vidicons. Their concept was first proposed by A. A. Chernyshev in 1925. The first commercial image orthicons and vidicons appeared in the United States in 1946-1950.

Speaking of the Soviet effort in the field of electronics, another direction in which outstanding discoveries and inventions were made was work on microwave devices and circuits. In 1932, D. A. Rozhansky came up with the idea of using velocity modulation in microwave devices. Following his suggestions, A. N. Arsenyeva and O. Heil built the first such devices to generate and amplify microwave oscillations. Later called drift-tube klystrons, they were also worked upon by R. Varian and S. Varian in the United States. In 1940, V. F. Kovalenko of the Soviet Union invented the simpler reflex klystron which is now widely used to generate and amplify microwave signals.

In 1938-1941, E. N. Daniltsev, V. K. Khokhlov, N. D. Devyatkov and M. D. Gurevich in the USSR designed disc-seal or planar-grid tubes for use in the UHF band. The active portions of the tube structure are in the form of planes or discs. Connections to the external resonators or cavities are made by means of metal rings attached to the disc electrodes. This principle was embodied in the metal-ceramic tubes made in Germany and in the lighthouse tubes that appeared at about the same time in the United States.

High power output in the microwave region is supplied by the magnetron, a version of the thermionic vacuum tube. The single-anode magnetron was initially investigated by A. W. Hull of the United States in 1921. His work was followed up by a large group of scientists in the Soviet Union (A. A. Slutsky, M. T. Grekhova,

D. S. Steinberg, V. I. Kalinin, S. A. Zusmanovsky, V. S. Lukoshkov, and S. Ya. Braude), Japan (H. Yagi and K. Okabe), France (L. Brillouin), Germany (E. Habann), and elsewhere. In the USSR, this type of tube as we know it today dates back, however, to 1936-1937 when V. F. Alexeyev and D. E. Malyarov developed what has come to be called the multicavity magnetron. In the United States, the first high-power microwave magnetron was built by J. T. Randall and H. A. H. Boot in 1940.

The 1930s and the later years saw a very rapid advance in semiconductor electronics. In the Soviet Union, especially valuable contributions in this field were made by a team of researchers under A. F. Ioffe in Leningrad. Their work covered physical processes in semiconductors, the effects of impurities on these processes, the thermoelectric and photoelectric properties of semiconductors, the rectification of alternating current by semiconductors, and many other issues. The theory of semiconductors formulated by the Ioffe school was later convincingly verified by experiments made in the Soviet Union.

Mention should also be made of B. I. Davydov who was the first to theorize on the rectification of alternating current at the metal-semiconductor junction. His theory was later elaborated by W. Schottky of Germany. Ya. I. Frenkel came out with a quantum theory of semiconductors, proposed the concept of movable vacancies in the crystal lattice of semiconductors (later called 'holes'), and formulated a theory explaining electron-hole pair generation.

Also in the 1930s, copper-oxide and selenium rectifiers were commercialized in the Soviet Union. Ya. I. Frenkel, L. D. Landau, B. I. Davydov and some others formulated a theory to explain the generation of an emf by illuminated semiconductors. A team under A. F. Ioffe built semiconductor thermo-emf batteries which later found a very broad field of application. Some of them could be put around the glass chimney of a kerosene lamp in order to generate electricity in an amount sufficient to run a portable radio transmitter-receiver. Today's space vehicles widely use as sources of electric power solar batteries which are assemblies of semiconductor thermocells.

In the 1940s, germanium and silicon diodes, semiconductor thermoresistors and photoresistors were commercialized in the Soviet Union. In 1948, the first type of transistor, called the point-contact transistor, was officially announced by V. Bardeen and W. H. Brattain of the United States. It was very much like the old familiar crystal detector but had two contacts known as cat whiskers that made contact on a small block of germanium. In the same year 1948, W. Shockley, also of the United States, invented what we know as the junction transistor, a semiconductor device which consists of a sandwich of alternate layers of *n*- and *p*-type germanium or silicon. In 1949, both the point-contact type and the junction type were put in production in the USSR.

In 1958-1960, a team under V. M. Tuchkevich of the USSR designed and commercialized high-power silicon-controlled rectifiers and thyristors. Ten years later, the population of these devices all over the USSR totalled around 10 million units capable of handling between them about 500 million kilowatts of power. In 1959, V. M. Wald-Perlov and A. S. Tager of the USSR invented the avalanche transit time diode adapted to generating microwave oscillations.

In 1972, V. M. Tuchkevich's team were awarded the Lenin Prize for heterojunction semiconductor devices, that is, devices using dissimilar semiconductor materials of opposite polarity. Their work had been greatly assisted by the contributions from Zh.I. Alferov of the Soviet Union.

Since then, many more discoveries and inventions have been made and embodied in practical devices. Some of them will be taken up in the text.

1-3 Requirements for Electronic Components

The various electron devices fall in the class of active electronic components because they can rectify, amplify, generate, and change the frequency of a. c. signals and perform other active processes. In contrast, there are passive components, such as resistors, capacitors, inductors, and transformers. Whatever the class of a given electronic component, there are some general requirements that they must all satisfy to be fit for use.

Component manufacturers, standards or specifications establish *ratings* – safe and limiting capabilities or conditions under which the components may be operated. These include nominal (or rated) voltages, currents, power dissipation, and the like. Since no component can be manufactured to have the precise rating, it is customary to specify also the respective *tolerances* usually expressed as percentages, for example, ±10%.

It is vitally important for electronic components to have ample *reliability*. This refers to the ability of any device, component, or circuit to perform a required function under stated conditions for a stated period of time. This may be expressed as a probability. 'Time' may be considered as distance, cycles, or other appropriate units. Reliability characteristics are those quantities used to express reliability in numerical terms. Electronic items are usually approached from the viewpoint of failure, service life, maintainability, and shelf life.

A *failure* refers to a complete or partial loss of the ability to perform a required function by a device, component, circuit, or any part or subsystem that can be separately tested.

Service life, or *lifetime*, is usually limited by the fact that some criterion (or criteria) of an item's performance degrades with time to a point where the item may no longer be used at all or it may be used in a less critical application, say, for educational purposes.

Maintainability refers to the ability of a device, component, or circuit to be kept in proper operating condition. Since any piece or equipment is less than 100% reliable, it is ultimately necessary to repair and maintain it. Also, in systems of any complexity, certain trimming adjustments are necessary at various stages of operation. Thus, a vital consideration in the design of the components and equipments of a system is the question of how quickly and easily a unit in that system can be adjusted or repaired. However, semiconductor devices and electron tubes fall in the class of *nonrecoverable* items, that is, those which cannot be brought back to a normal service status after a failure. For them, maintainability should be construed as the adaptability of an item to an easy check and replacement.

Shelf life refers to the length of time under specified conditions that a component retains its usability. It is usually found by a shelf test designed to measure the retention of serviceability after storage or transit under specified conditions of temperature.

Quantitatively, reliability may be expressed in terms of several criteria. Most often, this is done in terms of the *failure rate* symbolized by the Greek letter 'lambda' (λ). One of the definitions for the failure rate is as the ratio of the number n of like devices, components, etc., failing during an interval of time t, to the number N of devices, components, etc., operating at the start of the interval, multiplied by the time interval, or mathematically

$$\lambda = n/Nt \qquad \text{(I-1)}$$

If the time is in hours, the unit of λ will be the hour to the minus first power (that is, the reciprocal of the hour), h^{-1}. In other words, the failure rate may be defined as the fraction of components failing during one hour of operation. For example, if the number of components under test is $N = 1000$, if they operated for $t = 500$ h, and two components failed in the meantime, the failure rate will be

$$\lambda = 2/(1000 \times 500) = 4 \times 10^{-6}\ \mathrm{h}^{-1}$$

Or in words, the reliability is such that four components out of a million may fail during one hour of operation.

Failure is ordinarily classed according to cause, suddenness, and degree. Thus, failure can be *sudden* or *gradual*, that is, unanticipated by prior examination or anticipated. In the latter case, the primary cause is *ageing* or *wear-out*. It can also be *partial, complete,* or *intermittent*. A complete failure renders the failing component totally unfit for further use. In the case of a partial failure, the failing component may be used subject to certain limitations. There are various causes of failure: *misuse failure* is attributable to the application of stresses beyond the stated capabilities of the item; *inherent weakness failure* is attributable to a weakness inherent in the item itself when subjected to stresses within its stated capabilities, this weakness being due to either a poor design or to a poor workmanship.

A failure that is both sudden and complete is termed *catastrophic*; one that is both gradual and partial is called *degradation failure*.

The failure rate λ of an item varies during its lifetime (Fig. I-1). In the *early failure (*or *shakedown) period*, the failure rate starts from a relatively high value due to defects passed unnoticed during quality control or misuse during the early days of operation. This is what

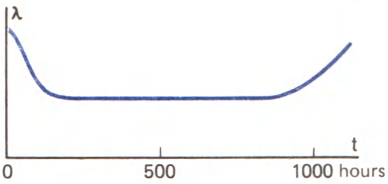


Fig. I-1
Failure rate-vs-operating age ('bathtub') curve

is known as the *infant mortality*, and it decreases rapidly. Next comes the period when the item is in useful operation. This period lies between the end of the shakedown period and the advent of wear-out. During this time, the failure rate of mature, well-designed equipment is at a low, nearly constant value. This period is the *useful life* of the item. With the advent of *wear-out*, the failure rate increases rapidly due to degradation processes. In the failure rate curve, sometimes called the '*bathtub*' *curve*, this interval is marked as the *wear-out failure period*. The reliability of electronic equipment degrades on aircraft and, especially, missiles. To avoid an early failure in use, a manufacturer will operate and test an item in a process known as *burn-in*; it stabilizes the item's characteristics. The period when an item is in useful operation is labelled on the failure rate curve as the *constant-failure rate period.*

It is vitally important for electronic devices, components, and equipments to have an adequate resistance to various exposures and a good parameter stability. The list of hostile exposures is topped by temperature. Electronic components must be *temperature-stable* – this means that their parameters should remain as constant with changes in temperature as practicable. In most cases, an item will be heated by the current that is flowing through it, by nearby components, and by the ambient air. This is the reason why measures should be taken to shield an item against heat uptake and to withdraw excess heat from it (by cooling or with the aid of a heat sink).

The temperature stability of a component is usually stated in terms of the respective temperature coefficient. For example, in the case of resistors one uses the temperature coefficient of resistance (TCR) which is defined as the incremental change in the resistance of a resistor as a result of a change in thermodynamic temperature by one degree and is expressed in kelvins to the minus first power:

$$\mathrm{TCR} = \Delta R/R\Delta t \qquad \text{(I-2)}$$

where ΔR is the change in the resistance of the resistor caused by a change in temperature by Δt.

For example, if $\mathrm{TCR} = 5 \times 10^{-4}\ \mathrm{K}^{-1}$, this means that heating a resistor by 1 K will change its resistance by 5×10^{-4} of its original value. If $R = 10\ \mathrm{k\Omega}$, this change will be 5 Ω per kelvin of temperature increase.

Heat resistance and *frost resistance* refer to the highest and the lowest temperature at which an item is still capable of performing its assigned function normally and there will be no failure.

Moisture resistance refers to the ability of an item to resist exposure to atmospheric moisture (humidity). Where there is a risk of water ingress in an electronic equipment, the latter must be made water resistant or water-proofed. Protection against humidity and water is provided by the use of protective films and other coatings and also by the encapsulation of components or even of a complete equipment.

Stability towards an elevated or a reduced pressure is important for electronic components that are likely to be used under such conditions. It is important to remember that the cooling of (the heat withdrawal from) an item is impaired as the ambient pressure goes down.

In some applications, it is important for electronic components to be *chemically stable.* This may be the case when an item is likely to be exposed to corrosive gases or fumes or when sea water is likely to get inside an item or an equipment.

If an equipment is to be used in a dust-laden atmosphere, such as may exist in a desert, the equipment itself and all of its components must be *dust-proofed.*

Radiation stability refers to the ability of an item to operate normally while being exposed to visible light or ionizing radiation. Unfortunately, some semiconductor devices cannot stand up properly to radiation.

An important consideration is the resistance to mechanical influences. This includes *shock resistance* and *vibration resistance*. The latter is especially important for electronic equipments carried on board ships, aircraft, and rockets.

Special mention should be made of *tropicalization* which refers to some form of design or treatment to combat the fungi that ruin electronic equipment in hot, humid tropical regions and also to repel attack by some dielectric-eating insects. Thus designed or treated, the items are termed tropicalized.

There are several special requirements that must be met by electronic devices, components, circuits, and equipments. Among other things, they must be able to perform normally in a desired *frequency range* and show an adequate *speed of response*. This is the reason why the data sheet of an electronic item will usually specify its operating frequency or its frequency limit.

As a rule, electronic components should preferably draw as little power from their power supplies as practicable (this is especially important for portable equipments) for their operation and also dissipate negligible power as heat.

Electronic elements are required to have a specified *dielectric strength* (also called electric strength, breakdown strength, and electric field strength). It is usually stated as the maximum voltage that the dielectric of a given item can withstand without rupturing. Sometimes, the respective maximum current or maximum power may be stated by a maker.

It is always desirable that an electronic item should be as small in size and as light in weight as achievable at the present state of the art. The reason is that today's electronic equipments pack each a very large number of components. This objective is currently achieved through *miniaturization* and *microminiaturization*. However, this approach poses a problem: the smaller an item, the lower its power rating, that is, the maximum power it can safely dissipate.

Importantly, electronic devices, components and circuits must lend themselves readily to a streamlined technology. This quality is usually referred to as *producibility* in the manufacturing industries. Also, they should preferably be made by a process that can easily be *mechanized* and *automated* because the huge multitude of electronic components turned out currently by their makers in amazing numbers cannot be fabricated with sufficient accuracy and, especially, repeatability by hand. There is an obvious advantage in the attempts by some countries to adopt common standards.

The cost of electronic components is a crucial economic factor, but they must of necessity be manufactured to the most stringent quality requirements, so their production cannot but be cost-intensive.

I-4 Semiconductor Devices in Electronics

In its early days and for several decades that followed, electronics relied almost exclusively on vacuum and gas-filled tubes. For some time past, however, semiconductor devices have come to be the basis of state-of-the-art electronics in nearly all of its divisions. This is the reason why we will begin our study by discussing semiconductor devices, and vacuum and gas-filled tubes will come in at a later time.

Crystal detectors which are radio-frequency semiconductor diodes have been in use in electronics since the advent of radio communication. Copper-oxide and selenium rectifiers have been used to rectify alternating current. For all their past records, however, the principle by which crystal detectors and rectifiers operate remained unclear for a long time.

As compared with vacuum tubes, semiconductor devices have several important advantages to offer, namely:

- They are light in weight and small in size.
- They need no heater (or filament) power.
- They have a higher reliability and a longer service life (tens of thousands of hours or even more).
- They are more robust mechanically (they readily stand up to vibration, impacts, and other mechanical factors).
- They have a far better efficiency because very little power is dissipated in semiconductor devices themselves.
- They need low voltages for their operation.
- They can be used in microelectronic circuits.
- They are far more cheaper to make.

However, they are prone to some limitations, namely:

- The items in a single batch of semiconductor devices may widely differ in their parameters and characteristics.
- Their behaviour and performance are strongly temperature-dependent.
- With time the properties and parameters of some semiconductor devices degrade (due to ageing).
- Inherent noise is sometimes greater than it is in vacuum tubes.
- Many transistor types cannot be operated at high frequencies.
- The input resistance of many transistors is only a fraction of that of vacuum tubes.
- The useful power output of transistors is so

far smaller than that supplied by tubes.

– The performance of most semiconductor devices is strongly degraded by exposure to radioactive emissions.

Extensive work is under way in all industrially advanced countries on improvements in semiconductor devices and on the use of novel, often unorthodox materials. Among the new additions to the electronic components on the market are semiconductor rectifiers for currents of thousands of amperes, transistors for frequencies of many hundred megahertz, and novel types of semiconductor devices capable of operation at frequencies running in to the gigahertz region.

Transistors are able to operate in nearly every type of electronic equipment using tubes, except some microwave equipments. As of this writing, transistors are used in amplifiers, receivers, transmitters, oscillators, TV receivers, measuring instruments, pulse circuits, computers, and many other applications.

Semiconductor devices offer their users huge savings in power and a chance to cut down the size and weight of their equipments many times. The minimum power needed to drive a vacuum tube is 0.1 W and more; for a transistor it may be as low as 1 μW, that is, 1/100 000th of its previous value.

In semiconductor integrated circuits (ICs), a silicon chip a few square millimetres in area can hold hundreds and even thousands of transistors. With such IC chips, one can readily build computers containing many millions of integrated components.

Transistors are the heart of miniature radio receivers and transmitters. A single flashlight battery is quite enough to sustain operation of such a set for many hours. Miniature radio components have specially been designed for use as companions for semiconductor devices, and together they have led to the advent of extremely small electronic equipment. For example, there are transceivers built into a handset and driven by the voice of the man speaking into the microphone. There are superminiature transistor radio transmitters enclosed with some other equipment into a capsule that can be swallowed by a patient and relay back data about his stomach and intestines.

Part One
Semiconductor Devices

Chapter One
Electric Conduction in Semiconductors

1-1 Electrons in Solids

It is proved in physics that electrons in a solid cannot possess just any arbitrary energy. Each electron can only have a particular discrete energy called an *energy level*.

The electrons closer to the atomic nucleus have lower energies, that is, they occupy lower energy levels. If we wish to move an electron away from the nucleus, we must overcome the mutual attraction between the electron and the nucleus, and this involves the expenditure of some energy. Therefore, the more distant electrons are more energetic, that is, they reside at higher energy levels.

When an electron moves from a higher to a lower energy level, it releases an amount of energy called a *quantum*; a quantum of electromagnetic radiation is a *photon*. If an atom absorbs one quantum of energy, the electron moves from a lower to a higher energy level. Thus, the energy carried by electrons can only change portionwise, that is, in quanta.

The distribution of electrons among the energy levels is usually shown on a diagram such as appears in Fig. 1-1. The horizontal lines drawn across the energy diagram represent each the energy W that an electron residing at that level has.

As has been proved by the band theory of solids, the energy levels form closely spaced groups called *energy bands* or, simply, *bands*. The electrons occupying the external shell of an atom fill a number of energy levels that form what is known as the *valence band*. The *valence electrons* take part in electrical and chemical processes. The lower energy levels are included into other electron-filled bands, but these bands (omitted in the energy diagram) do not contribute anything to the process of electric conduction.

Metals and semiconductors have a great number of electrons occupying the higher energy levels. These levels constitute the *conduction band*, and the electrons filling it are referred to as *conduction electrons*. They are moving at random, or roaming, from one atom to another. It is conduction electrons that are responsible for the high electrical conductivity of metals.

Atoms that give up their electrons to the conduction band may be viewed as positive ions. They make up an ordered arrangement called the *space lattice*, the *ion lattice*, or the *crystal lattice*. The state of this lattice corresponds to an equilibrium between the forces of interaction among its atoms, when the total energy of all particles in a specimen takes on a minimal value. Inside the lattice, conduction electrons are moving in a haphazard way.

Figure 1-1*a* shows an *energy level* (or *energy band*) *diagram* for a metal. It is important to

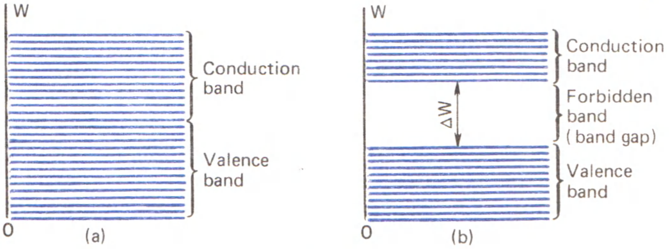


Fig. 1-1

Energy-level (or energy-band) diagram of (*a*) a metal and (*b*) a dielectric

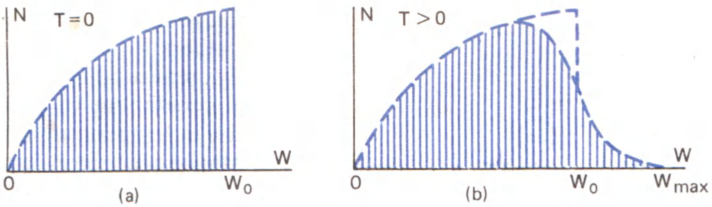


Fig. 1-2
Electron distribution by energy levels in a metal

stress that the actual pattern of energy bands is far more complicated, it has a very great number of energy levels, and the levels are distributed nonuniformly. If we write the maximum energy of electrons at absolute zero ($T = 0$) as W_0, the energy diagram may be drawn up as it appears in Fig. 1-2. The energy W is laid off horizontally, and the vertical segments represent the number N of electrons that have the energy shown (actually, there is a very great number of these vertical segments). The energy diagram in Fig. 1-2*a* corresponds to absolute zero. It shows that the number of electrons devoid of energy is zero. The higher the value of energy, the greater the number of electrons having it. The maximum number of electrons have an energy equal to W_0. The energy diagram in Fig. 1-2*b* is valid for a higher temperature. Now some electrons have an energy exceeding W_0 and a proportionately lesser number of electrons have an energy lower than W_0. The number of electrons having an energy in excess of W_0 decreases with increasing energy. The higher the temperature of the material, the higher the maximum energy W_{max}.

Figure 1-1*a* shows that in metals there is no gap between the conduction band and the valence band. Therefore, even at normal temperature a great number of electrons in metals have an energy sufficient for them to pass from the valence to the conduction band. Practically every atom of a metal donates at least one electron to the conduction band. Quite logically, the conduction electrons in metals are as numerous as the atoms.

The pattern of energy bands is different in dielectrics. In them, the conduction band is separated from the valence band by what is called a 'forbidden band', that is, one where no electrons can reside (Fig. 1-1*b*). The width, ΔW, of the forbidden band, or the difference in energy between the top level of the valence band and the bottom level of the conduction band, is a few electron-volts (eV). At normal temperature, the conduction band of a dielectric is populated by a very small number of electrons, and so the dielectric has a negligibly small electric conductivity. On heating, however, some electrons in the valence band acquire an additional energy and jump into the conduction band so that the dielectric displays a noticeable conductivity.

For semiconductors, the pattern of energy bands is similar to that shown in Fig. 1-1*b*, but the forbidden band is narrower than it is in dielectrics, being of the order of 1 eV in most cases. This is the reason why semiconductors behave as dielectrics at low temperatures while at normal temperature a substantial number of electrons jump from the valence to the conduction band. Electric conduction in semiconductors is discussed in more detail in the sections that follow.

At this writing, semiconductor devices are most often fabricated from germanium (Ge) and silicon (Si) both of which are four-valent substances. This means that the outer shells of a germanium or a silicon atom have four valence electrons. The lattice of germanium or silicon consists of atoms bound together by valence electrons. This type of linkage is known as the *covalent bond* and is shown in Fig. 1-3. As is seen, each pair of atoms is orbited by two valence electrons shown as small filled circles or dots. In a two-dimensional view of the lattice (Fig. 1-4), the covalent bonds are shown as straight lines, and the shared electrons as small filled circles (or they may be omitted altogether). As is seen, each atom of a pair contributes one electron to the shared pair that constitutes an ordinary chemical bond.

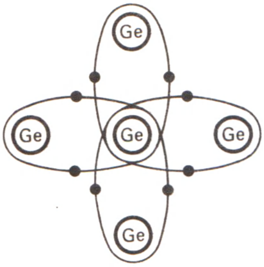


Fig. 1-3
Covalent bonding between germanium atoms

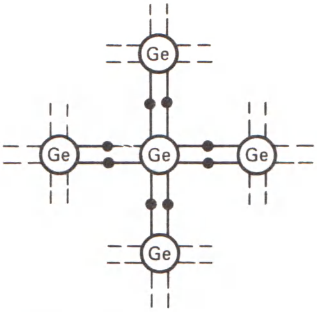


Fig. 1-4
Two-dimensional model of the germanium crystal lattice

1-2 Intrinsic Electron and Hole Conduction. Drift Current

In terms of electric conductivity,* semiconductors stand midway between conductors and dielectrics.

At $T = 300$ K, the electric conductivity of conductors is 10^4-10^6 S cm^{-1}. (As will be recalled 1 S cm^{-1} is the electric conductivity of 1 cm^3 of a material.) The figure for dielectrics is less than 10^{-10} S cm^{-1}, and for semiconductors it ranges from 10^{-10} to 10^4 S cm^{-1}. As is seen, the electric conductivity of semiconductors extends across a very broad range. Most substances occurring in nature are semiconductors. At this writing, semiconductors used for commercial purposes include primarily germanium and silicon, and also gallium arsenide (GaAs), indium antimonide (InSb), indium phosphide (InP), and some others.

Typically, semiconductors have a negative temperature coefficient of resistance. In other words, a rise in temperature brings about a fall in the resistance of semiconductors, and not the other way around as happens with most solid conductors. Also, the resistance of semiconductors heavily depends on their impurity content and exposure to external factors, such as light, an electric field, an ionizing radiation, etc.

Semiconductor diodes and transistors depend for their operation on the fact that two types of electric conduction exist in semiconductor materials. Like metals, they show *electron conduction*, that is, the flow of current due to the motion of conduction electrons. At ordinary operating temperatures, any semiconductor always has conduction electrons which are only loosely tied to the atomic nuclei and are in a random thermal motion among the lattice atoms. When exposed to a potential difference, these roaming electrons can constitute an ordered flow while retaining their roaming behaviour. This additional flow is an electric current.

The other type of conduction displayed by semiconductors but nonexistent in metals is *hole conduction*. Since it is a distinction of semiconductors, hole conduction deserves a more detailed discussion.

In a semiconductor atom, thermal or other factors may cause one of the valence electrons most distant from the nucleus to jump into the conduction band. As a result, the atom will acquire a positive charge numerically equal to that on the electron. This atom may be called a positive ion. It should be borne in mind, however, that in the case of the true ionic conduction, such as is observed in electrolytes, the current is constituted by the flow of ions (the very word 'ion' means 'a traveller'). In hole conduction, an entirely different mechanism is involved – the crystal lattice of semiconductors is sufficiently strong, so its ions remain stationary rather than move.

The vacancy left in the valence band of a semiconductor due to an electron being lost from the band by, say, thermal excitation has come to be known as a *hole*. Holes behave like positive charge carriers.

The way a hole is produced can be seen in Fig. 1-5 which is the already familiar two-di-

* Not to be confused with 'electric conduction', which is the ability of a material to transmit electricity. Electric conductivity is the numerical measure of electric conduction.

mensional model of a semiconductor. On receiving an additional energy, one of the electrons contributing to a covalent bond becomes a conduction electron, that is, a charge carrier which is free to move about in the crystal lattice. As it does so, it leaves behind a vacancy, that is, a hole shown as an open circle in the figure.

Under the influence of an applied potential difference (or electric field), a hole is caused to be propagated through the lattice, which is equivalent to the motion of a positive charge. This process is depicted in Fig. 1-6 which shows several atoms of a semiconductor material at several instants. Let, at the initial instant (Fig. 1-6*a*), a hole be produced in the atom on the extreme left (*1*) due to its loss of an electron. The electron-deficient atom (shown shaded) is positively charged and can attract an electron from one of its neighbours (*2*). When an electric field (or a potential difference) is applied to the material, the field tends to move electrons from the negative towards the positive potential. Therefore, during the next instant (Fig. 1-6*b*) one electron from atom *2* jumps into atom *1* and fills the vacant hole, while leaving a new vacancy, or hole, behind in atom *2*. Then an electron from atom *3* will jump into atom *2* to fill its vacant hole, thereby causing a further hole to appear in atom *3* (Fig. 1-6*c*), and so on. This chain of events will go on until the hole has moved from the atom on the extreme left to the atom on the extreme right. In this way, the positive charge originally produced in atom *1* will have moved to atom *6* (Fig. 1-6*f*).

Thus, as we have seen, the electric current produced in the case of hole conduction is likewise due to the motion of electrons, but their motion is more limited than in the case of electron conduction. An electron can move from an atom only to its neighbour. Also, the effective movement of the hole due to a process of continuous exchange is in the direction of the positive field, that is, opposite to the direction of electron movement.

The situation we have just described may be likened to a concert-hall with rows of chairs occupied by spectators. Suppose that a spectator in the first row leaves his seat and a spectator from the row next behind him fills the vacant place. In turn, a spectator in the third row moves forward and occupies the vacant seat in the second row. Finally, a spectator from the last row moves forward to take up the vacant seat in the last-but-one row. Thus, the vacant seat originally left in the first row has finally moved to the last row. In this example, the spectators act similarly to the electrons and the consecutively vacated seats similarly to the holes in our previous example. Thus, the seats (that is, 'holes') remained stationary, and only the spectators (that is 'electrons') move in succession.

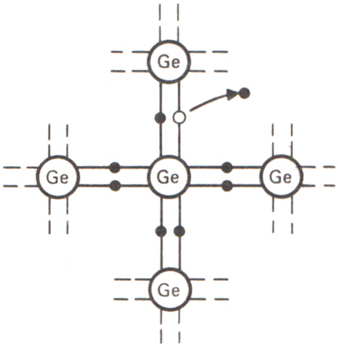


Fig. 1-5
Electron-hole pair generation

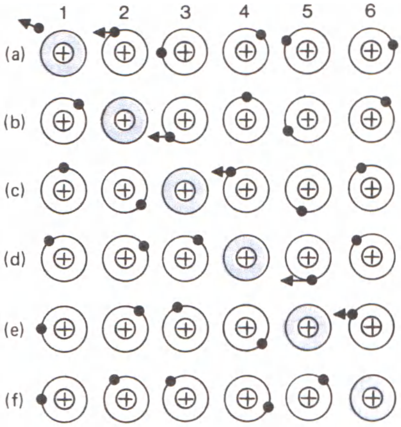


Fig. 1-6
Hole conduction

The best way to explain electric conduction in semiconductors is by reference to their energy band diagram (Fig. 1-7). As we know, the width of the forbidden band (band gap) in semiconductors is relatively small (being 0.72 eV for germanium and 1.12 eV for silicon). At absolute zero, a semiconductor free from impurities is a dielectric – it has neither conduction electrons

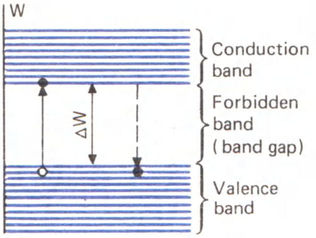


Fig. 1-7

Pattern of energy levels in a semiconductor

nor holes. As its temperature is raised, however, the semiconductor acquires an ever greater electric conductivity because heating imparts an additional energy to the electrons in the valence band, and an ever greater number of them cross the forbidden band and jump from the valence band into the conduction band. This is illustrated in Fig. 1-7 by a solid arrow. That is how conduction electrons are produced and electron conduction takes place. Each electron moving into the conduction band leaves behind in the valence band a vacancy, or a hole. In this way, the valence band acquires conduction holes, and their number is equal to that of electrons that have moved into the conduction band. Thus, electron conduction is accompanied by hole conduction.

Electrons and holes that can move and bring about electric conduction are called *mobile charge carriers* or simply *charge carriers* (or still simpler, *carriers*). It is customary to say that thermal excitation results in the *generation of electron-hole pairs*. Electron-hole pair generation may likewise be brought about by light, an electric field, an ionizing radiation, etc.

Both conduction electrons and holes move in a random fashion, and this is responsible for a process which is the reverse of electron-hole pair generation. Conduction electrons eventually reoccupy vacancies in the valence band, that is, recombine with holes. Quite aptly, this occurrence is referred to as *electron-hole pair recombination*. During this process, an electron moves from the conduction band into the valence band as shown by the dashed arrow in Fig. 1-7. Electron-hole pair generation and recombination occur at the same time always. Recombination limits the increase in the number of electron-hole pairs, so for each particular temperature of the material there is a certain definite number of each species. In this way, their populations are in a state of dynamic equilibrium. To state this a bit differently, ever new pairs of electrons and holes are generated all the time, and those produced earlier recombine as continuously.

A semiconductor free from impurities is called an *intrinsic*, or *i-type*, *semiconductor*. It possesses *intrinsic electric conduction* which, as has been shown, is a combination of electron conduction and hole conduction. Importantly, although an intrinsic semiconductor has equal concentrations of electrons and holes under conditions of thermal equilibrium, electron conduction is predominant. This is because electrons have a greater mobility as compared with holes. It is easy to understand why this is so: in the case of hole conduction, the electrons are more limited (less free) to move about than in the case of electron conduction.

The electric conductivity of a semiconductor depends on its *carrier concentration* which is defined as the number of charge carriers per unit volume, in practice usually quoted as a number per cubic centimetre.* One way to symbolize the carrier concentration is with the letters n (for 'negative') in the case of electron concentration and p (for 'positive') in the case of hole concentration. Obviously, in an intrinsic semiconductor

$$n_{\mathrm{i}} = p_{\mathrm{i}}$$

where the subscript "i" indicates that we refer to an intrinsic semiconductor.

One cubic centimetre of a metal or semiconductor specimen has a number N of atoms of the order of 10^{22}. At a temperature close to 20°C, the (approximate) carrier concentration for pure germanium is

$$n_{\mathrm{i}} = p_{\mathrm{i}} = 10^{13}\ \mathrm{cm}^{-3}$$

and for silicon,

$$n_{\mathrm{i}} = p_{\mathrm{i}} = 10^{10}\ \mathrm{cm}^{-3}$$

Thus, at room temperature the ratio of mobile carriers to the total number of atoms in an intrinsic semiconductor is about $10^{-7}\%$ for germanium and about $10^{-10}\%$ for silicon. In metals, the number of conduction electrons is no less than that of atoms ($n \geqslant N$). Therefore, the electric conductivity of semiconductors is a few millionths or even a few thousand-millionths of

* Some authors term this quantity as the carrier density.– *Translator's note.*

that of metals. For example, the resistivity at room temperature is $0.017 \times 10^{-4}\,\Omega$ cm for copper, about 50 Ω cm for germanium, and about 100 000 Ω cm for silicon (1 Ω cm is the resistance of one cubic centimetre of a material).

If no voltage is applied to a semiconductor specimen, its conduction electrons and holes will be in a chaotic thermal motion, and there will of course be no net flow of current. When a potential difference is applied to a semiconductor specimen, it sets up an electric field which accelerates the electrons and holes and imparts them some translational motion which constitutes what is known as the conduction current.

An alternative name for the motion of charge carriers under the influence of an applied field is *drift*, and the associated current is referred to as the *drift current*, i_{dr}. The total conduction current is the sum of the electron drift current $i_{n,\mathrm{dr}}$ and the hole drift current $i_{p,\mathrm{dr}}$:

$$i_{\mathrm{dr}} = i_{n,\mathrm{dr}} + i_{p,\mathrm{dr}} \tag{1-1}$$

Although electrons and holes move in opposite directions, the two currents are added together because the motion of holes is in effect the motion of electrons. For example, in an intrinsic semiconductor $i_{n,\mathrm{dr}} = 6$ mA, and $i_{p,\mathrm{dr}} = 3$ mA, and so the total or net conduction current is $i_{\mathrm{dr}} = 9$ mA.

In order to establish the factors that affect the net current, it is convenient to consider the current density rather than the current itself. Obviously, the drift (net) current density J_{dr} is the algebraic sum of the electron and hole current densities

$$J_{\mathrm{dr}} = J_{n,\mathrm{dr}} + J_{p,\mathrm{dr}} \tag{1-2}$$

Since current density is defined as the ratio of the current to the cross-sectional area of the current-carrying medium per second, we may write the electron current density as

$$J_{n,\mathrm{dr}} = n_{\mathrm{i}} e u_n \tag{1-3}$$

where n_{i} is the concentration of electrons, e is the charge on an electron, and u_n is the average velocity at which electrons move translationally under the influence of an applied field.

It is important to remember that the average velocity takes into account the random thermal motion of electrons during which they collide with the atoms of the lattice many times. During the time interval from one collision to another each electron is accelerated by the field, and so u_n is proportional to the electric field strength E:

$$u_n = \mu_n E \tag{1-4}$$

where μ_n is a proportionality coefficient called the *electron mobility*. Its meaning is easy to grasp if, using Eq. (1-4), we write

$$\mu_n = u_n / E \tag{1-5}$$

As follows from Eq. (1-5), when $E = 1$, the electron mobility μ_n is equal to the average velocity of electrons as they move under the influence of an applied electric field of unity strength. If we express the velocity in centimetres per second and the field strength in volts per centimetre, the unit of carrier mobility will be

$$(\mathrm{cm\ s^{-1}})/(\mathrm{V\ cm^{-1}}) = \mathrm{cm^2\ V^{-1}\ s^{-1}}$$

For example, at room temperature the electron mobility for pure germanium is $3600\ \mathrm{cm^2\,V^{-1}\,s^{-1}}$. In other words, an electric field whose intensity is $1\ \mathrm{V\,cm^{-1}}$ will cause the conduction electrons in a specimen of pure germanium to move in the direction of the field with an average velocity of $3600\ \mathrm{cm\ s^{-1}}$. Electron mobility is different for different semiconductors and tends to decrease with rising temperature because the electrons collide with the atoms of the crystal lattice far more often.

On expressing the velocity in Eq. (1-3) in terms of $\mu_n E$, we get

$$J_{n,\mathrm{dr}} = n_{\mathrm{i}} e \mu_n E \tag{1-6}$$

The product $n_{\mathrm{i}} e \mu_n$ in Eq. (1-6) is the *electron conductivity* whose symbol is σ_n. Hence,

$$J_{n,\mathrm{dr}} = \sigma_n E \tag{1-7}$$

The relations and the reasoning given above may be extended to include conduction holes as well. Then, the hole current density will be given by

$$J_{p,\,\mathrm{dr}} = p_{\mathrm{i}} e \mu_p E \tag{1-8}$$

where the product $p_{\mathrm{i}} e \mu_p$ is the hole conductivity σ_p.

Hence, the total drift (net) current in an intrinsic semiconductor is

$$J_{\mathrm{dr}} = n_{\mathrm{i}} e \mu_n E + p_{\mathrm{i}} e \mu_p E = (\sigma_n + \sigma_p) E \tag{1-9}$$

and the total conductivity,

$$\sigma = \sigma_n + \sigma_p = n_{\mathrm{i}} e (\mu_n + \mu_p) \tag{1-10}$$

Thus the conductivity of a semiconductor is a function of both the carrier concentration and the carrier mobility. In semiconductors, a rise in temperature promotes electron-hole pair generation, and the mobile carrier concentration builds up faster than their mobility decreases. As a result, a rise in the temperature of a semiconductor specimen leads to an increase in its conductivity. By way of comparison, it may be noted that in metals the conduction electron concentration is almost independent of temperature, so a rise in the temperature of a metal specimen leads to a fall in its conductivity due to a decrease in the electron mobility.

It is also to be recalled that $\mu_p < \mu_n$ always, and so $\sigma_p < \sigma_n$. For example, at room temperature $\mu_n = 3600$ cm^2 V^{-1} s^{-1} and $\mu_p = 1820$ cm^2 V^{-1} s^{-1} for germanium, and $\mu_n = 1300$ cm^2 V^{-1} s^{-1} and $\mu_p = 460$ cm^2 V^{-1} s^{-1} for silicon.

1-3 Extrinsic Conduction

If a semiconductor contains other substances as impurities, it will display what may be called *extrinsic conduction* in addition to its intrinsic conduction. It depends on the type of impurity (or impurities) present, and may be electron or hole conduction. For example, if we *dope* (as it is said) pure germanium which has four electrons in its valence band with a controlled amount of a *donor* element having five electrons in its valence band, such as antimony (Sb), arsenic (As), or phosphorus (P), under proper physical conditions, atoms of the impurity element will take the places of germanium atoms in their crystal lattice, and four of the five valence electrons will perform in the manner of the four germanium valence electrons they have displaced. But the fifth valence electron of the impurity atom will be free to move on and form conduction current. Antimony, arsenic, phosphorus and other elements with five valence electrons are called donors because they donate electrons to the host crystal. On giving up their fifth valence electrons, the donor atoms become positively charged. How the above process takes places in the case of antimony as the donor impurity and a germanium crystal as the host is shown in the two-dimensional lattice pattern of Fig. 1-8.

Semiconductors showing a predominance of electron conduction are called *n-type semiconductors*. An energy band diagram for an *n*-type semiconductor is shown in Fig. 1-9. The energy levels of the donor atoms are situated only slightly below the conduction band of the host material. Therefore, one electron from each donor atom can readily jump into the conduction band, and an additional number of electrons, equal to that of donor atoms, is added to that band. No holes are produced in the donor atoms as a result of this process.

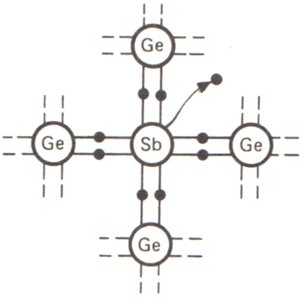


Fig. 1-8
Mechanism of extrinsic electron conduction

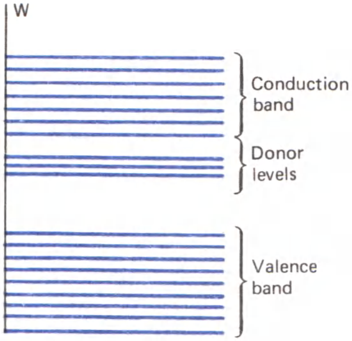


Fig. 1-9
Energy-band diagram of an *n*-type semiconductor

Now consider the elements boron (B), indium (In), and aluminium (Al). Each of them has only three electrons in their respective valence band. When controlled amounts of any one of these elements are added to highly refined germanium (Ge), their atoms will displace germanium atoms in their crystal lattice. In this case there will be one incomplete bond on a neighbouring germanium atom. The missing electron constitutes a hole which may be neutralized electrically by an electron jumping into it from nearby bond and thus effectively moving the hole to a new position in the germanium specimen. Boron, indium, aluminium, and other impurity elements with three valence electrons are called *acceptors* because they take on electrons from the crystal instead of donating them as do arsenic and antimony. On capturing these electrons, the acceptor atoms are charged negatively. The process we have just described is diagrammatically shown in Fig. 1-10.

Semiconductors showing a predominance of hole conduction are called *p-type semiconductors* (Fig. 1-11). The energy levels of the acceptor atoms are located only slightly above the valence band. The electrons from the valence band where holes are thus produced can readily jump into these levels.

Semiconductor devices are mostly made of semiconductor materials containing donor or acceptor impurities, and they are called *extrinsic semiconductors*. In such semiconductors all of the impurity atoms contribute to extrinsic conduction at normal operating temperatures by donating or accepting an electron each as the case may be.

For extrinsic conduction to exceed intrinsic conduction, the donor atom concentration N_d or the acceptor atom concentration N_a should exceed the intrinsic carrier concentration $n_i = p_i$. In the practical manufacture of extrinsic semiconductors, N_a or N_d is always many times n_i or p_i. For example, in the case of germanium which has $n_i = p_i = 10^{13}$ cm^{-3} at room temperature, N_d and N_a may range anywhere between 10^{15} and 10^{18} cm^{-3}, which is 10^2 to 10^5 times the intrinsic carrier concentration. In our subsequent discussion, we will give examples for germanium at room temperature.

Thus, conduction in *n*-type germanium at ordinary temperature is by means of electrons supplied by the donor impurity. These excess electrons are spoken of as *majority carriers*. Conduction in *p*-type germanium at ordinary temperatures is by means of holes that are created when an acceptor impurity is added. Now the 'doped' germanium has an excess of these 'missing negative charges' in its atoms. Holes are thus in the majority, and in this situation they are called the majority carriers. The remaining carriers of opposite sign in each case are called *minority carriers*.

If $N_d \gg n_i$, we may neglect the intrinsic carrier concentration (that is, that of electrons), and then $n_n^* \approx N_d$. Taking *n*-type germanium as an example, n_n may be about 10^{-16} cm^{-3}. Clearly, as compared with this figure, one need not consider the intrinsic carrier concentration which is $n_i = 10^{13}$ cm^{-1}, or by a factor of 1000 smaller.

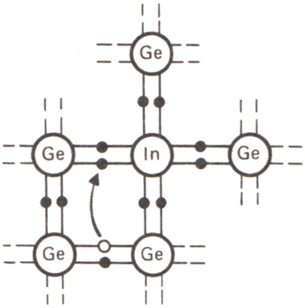


Fig. 1-10

Mechanism of extrinsic hole conduction

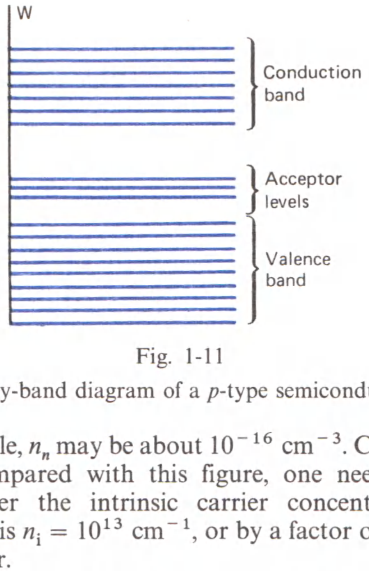


Fig. 1-11

Energy-band diagram of a *p*-type semiconductor

The minority carrier concentration in an extrinsic semiconductor decreases in the same proportion as the majority carrier concentration increases. Thus, if *i*-type germanium has $n_i = p_i = 10^{13}$ cm^{-3}, and doping it with a donor impurity increases the figure 1000-fold to $n_n = 10^{16}$ cm^{-3}, the minority carrier (hole) concentration will fall to 1/1000th of its previous value to become $p_n = 10^{10}$ cm^{-3}, which is one millionth of the majority carrier concentration. The explanation is that when the conduction electron concentration is increased 1000-fold due to the contribution from donor atoms, the lower energy levels in the conduction band are filled full, and electrons from the valence band can now jump only to the higher energy levels of the conduction band. For this 'jump' to occur,

* Here and elsewhere the subscripts *n* and *p* refer, respectively, to *n*- and *p*-semiconductors.

however, the electrons must have a greater energy than they do in an intrinsic semiconductor, and so much fewer electrons are capable of doing this. The number of holes in the valence band then decreases in about the same sizeable proportion. It has been found that n-type extrinsic semiconductors always satisfy the following equality:

$$n_n p_n = n_i p_i = n_i^2 = p_i^2 \tag{1-11}$$

In our example,

$$10^{16} \times 10^{10} = (10^{13})^2 = 10^{26}$$

Everything said about n-type semiconductors fully applies to p-type semiconductors. In them, $N_a \gg p_i$, and we may deem that $p_p \approx N_a$. For example, in the case of p-type germanium the figures may be $p_p = 10^{16}$ and $n_p = 10^{10}$ cm^{-3}. For p-type semiconductors, the equality

$$p_p n_p = n_i p_i = n_i^2 = p_i^2 \tag{1-12}$$

also holds always.

As we have seen, even minute amounts of an impurity can drastically change both the type of conduction and the magnitude of conductivity of a semiconductor. Indeed, with as many as 4.4×10^{22} germanium atoms enclosed in every cubic centimetre of the material, an impurity concentration of 10^{16} cm^{-3} will amount to adding only one impurity atom to the four-odd million germanium atoms. That is, the impurity will account for as little as 10^{-4}% of the total material. Nevertheless, the majority carrier concentration is increased 1000-fold and the conductivity is improved in the same proportion.

The manufacture of semiconductors carrying such negligible and closely controlled amounts of an impurity is a very sophisticated process. On top of that, the host material must be extremely pure – for germanium the tolerance on any unwanted impurities is not over 10^{-8}% which works out to not more than one atom per 10 thousand million germanium atoms. In silicon, the unwanted impurities may be present in an amount not exceeding 10^{-11}% which is a still tighter tolerance.

The electric conductivity of extrinsic semiconductors is determined in the same way as for intrinsic semiconductors. If we neglect the conductivity due to the minority carriers, then we may write for n-type and p-type semiconductors,

$$\sigma_n = n_n e \mu_n \text{ and } \sigma_p = p_p e \mu_p \tag{1-13}$$

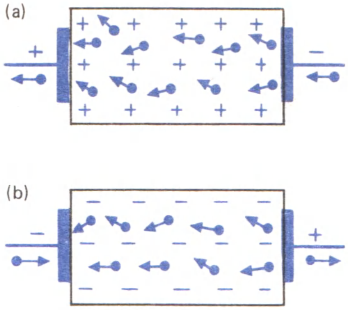


Fig. 1-12

Current in (*a*) an n-type semiconductor and (*b*) a p-type semiconductor

Consider the flow of current through each type of semiconductor, while neglecting the current due to the minority carriers for simplicity. As before, Fig. 1-12 shows holes as open (unshaded) circles, and electrons as filled circles (or dots). The "+" and "−" signs indicate the polarity of charge on the atoms of the crystal lattice. When an emf is applied from an external source to an n-type semiconductor, it causes conduction electrons to move both in the wires connecting the specimen to the emf source and in the specimen itself. In a p-type semiconductor an applied emf will likewise cause electrons to move in the connecting wires, but the current in the semiconductor should be regarded as the motion of holes. The electrons coming from the negative side fill the holes, and the positive side receives the electrons arriving there from the adjacent regions of the specimen where holes are thus produced, and they move from the right-hand to the left-hand side of the specimen (in the picture).

In electrical engineering, it is customary to think that an electric current flows from a "+" terminal to a "−" terminal. When considering electron devices, it is convenient to consider the actual direction of current flow – from the "−" to the "+" terminal. We will indicate this direction by an arrow with a bold dot at the start, and the assumed (or conventional) direction by an arrow without a dot.

1-4 Carrier Diffusion in Semiconductors

The carriers present in a semiconductor move by one of two mechanisms, drift or diffusion. Drift, as we have just seen, is the motion caused by the presence of an electric field – due to the field, the carriers acquire a directed component of motion. The sum of the directed components of drift constitute what we have called the drift current in the specimen. But carriers also move by the process of *diffusion*, giving rise to the *diffusion current*. Its cause is the difference in carrier concentration and not in potential between different parts of a specimen.

If the carriers in a specimen are distributed uniformly, we have an *equilibrium carrier concentration*. Some external factors may upset this equilibrium concentration so that it is higher in one region and lower in some other (a *nonequilibrium concentration*). For example, if we throw a beam of light on some part of a semiconductor specimen, more electron-hole pairs will be generated there, and this will produce what is known as an *excess concentration*.

Since carriers always have a kinetic energy of their own, they always tend to move from regions of high concentration to regions of low concentration until it is the same throughout the specimen.

Diffusion is likewise displayed by species other than mobile charge carriers. It is a proven fact, for example, that molecules do diffuse in many substances. Whatever the kind of species involved, however, the cause of diffusion is always the same - a difference in species concentration between various regions in a sample, and the diffusion itself is effected at the expense of the energy that the species involved possess due to thermal motion or excitation.

Similarly to the conduction current, the diffusion current, symbolized as i_{dif}, may be constituted by electrons or by holes. The respective densities are given by

$$J_{n\,\text{dif}} = eD_n\Delta n/\Delta x$$

and

$$J_{p\,\text{dif}} = -eD_p\Delta p/\Delta x \qquad (1\text{-}14)$$

where the terms $\Delta n/\Delta x$ and $\Delta p/\Delta x$ are referred to as the *concentration gradients*, and the quantities D_n and D_p as the *diffusion coefficients*.

The concentration gradient is a measure of the difference in carrier concentration (free electrons or holes) from point to point in a semiconductor per unit length. The greater the change in electron or hole concentration, Δn or Δp, over the distance Δx, the greater the diffusion current. No diffusion current exists when $\Delta n = 0$ or $\Delta p = 0$.

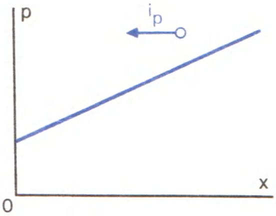


Fig. 1-13

Motion of holes in the presence of a concentration difference

The diffusion coefficient is a measure of the rate at which the process of diffusion occurs. It is proportional to the carrier mobility, differs from one material to another, and is a function of temperature, or mathematically

$$D = \mu kT/e$$

where μ is the carrier mobility, k is the Boltzmann constant, T is the thermodynamic temperature, and e is the electron charge. It is expressed in square centimetres per second. The diffusion coefficient of electrons is always greater than that of holes. Taking germanium at room temperature, it is $D_n = 98$ cm^2 s^{-1} and $D_p = 47$ cm^2 s^{-1}. For silicon, the respective figures are $D_n = 34$ cm^2 s^{-1} and $D_p = 12$ cm^2 s^{-1}.

The "–" sign in the equation defining the diffusion current density for holes indicates that the hole current is flowing towards a region of lower hole concentration. This is explained in Fig. 1-13 which shows that when the hole concentration builds up with increasing x, the holes are moving in a direction which is opposite to the positive x-axis. Hence, the hole current must be taken as negative.

If we let some external factor produce an excess carrier concentration in some region of a semiconductor and then remove this factor, the excess carriers will recombine and propagate by diffusion to other regions of the specimen. The excess concentration will decrease exponentially, as shown in the plot of Fig. 1-14 for the electron concentration. The time during which the excess concentration is reduced by a factor of 2.7, that is, falls to 0.37 of its original value, n_0, is called the *lifetime of nonequilibrium carriers*, τ_n. It describes how fast the excess concentration of carriers decreases.

Nonequilibrium carriers recombine both inside the semiconductor and on its surface, and the recombination rate strongly depends on the amount and kind of impurities present and the surface conditions. The value of τ_n for germanium and silicon may range from a fraction of a microsecond to hundreds of microseconds or even longer, depending on the circumstances.

When nonequilibrium carriers (say, electrons) propagate through a specimen by diffusion, their concentration likewise decreases with distance exponentially due to recombination (Fig. 1-15). The distance L_n over which the excess concentration of nonequilibrium (usually, minority) carriers is reduced by a factor of 2.7, that is, falls to 0.37 of its original value n_0, is called the *diffusion length*. It describes how the excess concentration decreases with distance. In more rigorous terms, it is defined as the average distance that the minority carriers move in a homogeneous semiconductor between generation and recombination.

Thus, the excess concentration decreases both with time and distance, and so the lifetime τ_n and the diffusion distance L_n are interconnected by a relation of the form

$$L_n = (D_n \tau_n)^{1/2} \tag{1-15}$$

All we have said about excess electrons fully applies to excess holes, but τ_p and L_p take on different values, of course.

Conduction current and diffusion current, electron-hole pair generation and recombination, changes in the excess carrier concentration with time and distance do not exhaust the multitude of complex processes that take place in semiconductors, but they are most important and their knowledge permits a proper insight into the workings of semiconductor devices.

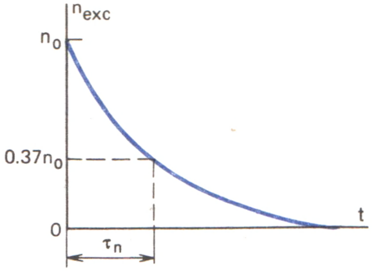


Fig. 1-14
Time variations in excess charge concentration

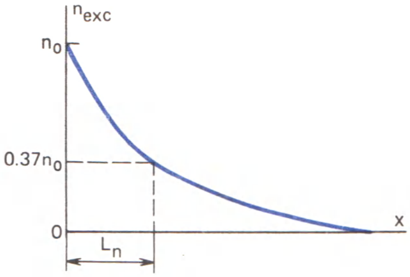


Fig. 1-15
Spatial variations in excess charge concentration

Chapter Two

P-N and Metal-Semiconductor Junctions

2-1 A *P-N* Junction with No External Voltage Applied

The term *junction* in our case refers to the boundary, or the region of transition, between *n*-type and *p*-type semiconductor materials; hence the name a *p-n* junction. A *p-n* junction has an unsymmetrical conductivity or, which is the same, it has a nonlinear resistance. Most semiconductor devices (diodes, transistors, etc.) depend for their operation on the properties of one or several *p-n* junctions. Let us take a closer look at what happens in a *p-n* junction.

We assume that no external voltage is applied across the *p-n* junction (Fig. 2-1). Because the carriers in each semiconductor move about in a random manner due to thermal excitation (that is, they have their own velocities), they diffuse from one semiconductor into the other. As with any form of diffusion, such as in gases or liquids, the carriers move from a region of high concentration to a region of low concentration. For this reason, electrons diffuse from the *n*-type semiconductor where their concentration is high into the *p*-type semiconductor where their concentration is low. Conversely, holes diffuse

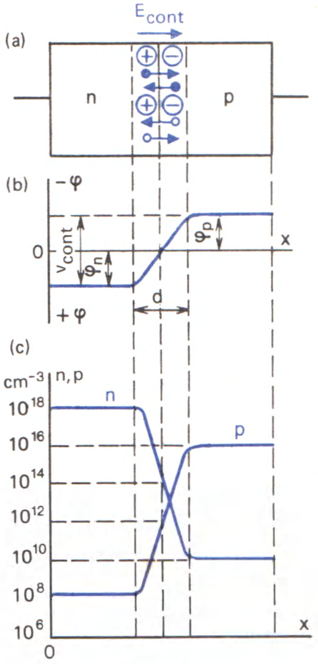


Fig. 2-1

P-N junction with no external bias applied

from the *p*-type semiconductor where their concentration is high into the *n*-type semiconductor where their concentration is low. This process is shown by the arrows in Fig. 2-1*a*. The circles labelled with the "+" and "−" signs represent the donor and acceptor impurity atoms charged positively and negatively, respectively.

As a result of the diffusion process described above, charges of opposite sign are produced on each side of the boundary between the *n*-type and *p*-type material, a positive charge in the *n*-region and a negative charge in the *p*-region. The positive charge in the *n*-region is mainly formed by the positively charged atoms of the donor impurity and, in part, by the holes that have come across the junction. The negative charge in the *p*-region is mainly formed by the negatively charged atoms of the acceptor impurity and, in part, by the electrons that have come across the boundary. For simplicity, Fig. 2-1*a* only shows the carriers and the impurity atoms in the transition region.

What is known as the *contact potential difference*

$$v_{cont} = \varphi_n - \varphi_p$$

is produced between the two charges which give rise to an electric field (field intensity vector E_{cont}). Figure 2-1*b* shows a potential diagram for the *p-n* junction that we have taken as an example, assuming that no external voltage is applied across the junction. In this diagram which shows the potential distribution along the *x*-axis perpendicular to the *p-n* junction, the boundary layer is assumed to be at zero potential. In fact, we may have taken either the *n*-region or the *p*-region to be at zero potential. The diagram in Fig. 2-1 and in the subsequent figures is shown on an exaggerated scale for ease of visualization. Actually, the *p-n* junction is extremely thin in comparison with the *n*- or *p*-region.

Importantly, space charges of opposite signs are produced near the boundary between the *n*- and *p*-regions, while the positive potential φ_n or the negative potential φ_p is the same throughout the *n*-region or *p*-region, respectively. If different potentials existed in different parts of, say, the *n*- or *p*-region, this would produce a current which would finally equalize the potential throughout the respective region all the same. It should be remembered that charge and potential have a different physical meaning. A region at some potential need not necessarily possess a charge.

Thus, each kind of charge finds itself surrounded by charges of opposite sign, and recombination takes place. Figure 2-1*a* shows the junction after diffusion and recombination have happened. The boundary region soon becomes devoid of holes on the *p*-side and of electrons on the *n*-side. The result is an accumulation of positive charges at the border on the *n*-side and of negative charges on the *p*-side of the border. These are called *bound charges* by some authors. They are positive and negative ions that cannot move about. Two equal and opposite charges separated a small distance are called a *dipole*, and in this situation there is what is called a *dipole layer*. Such a layer is present in all *p-n* junctions.

Free electrons on the *n*-side cannot go over to the *p*-side because of the opposing forces of its negative ions, and free holes on the *p*-side cannot go over to the *n*-side because of the opposing forces of its positive ions.

As is seen, a *potential barrier* is produced in the *p-n* junction, preventing any further diffusion of carriers into the opposite regions. Figure 2-1*b* shows a potential barrier for electrons which tend to move by diffusion from left to right (from the n- to the p-region). If we lay off the positive potential upwards, we can plot an image of a similar potential barrier for holes tending to diffuse from right to left (from the p- into the n-region).

The height of this barrier is equal to the contact potential difference and usually is a few tenths of a volt. The higher the impurity concentration, the greater the majority carrier density, and the greater the number of majority carriers that are capable of diffusing across the boundary. The space charge density increases, and the contact potential difference v_{cont} goes up, which is another way of saying that the potential barrier builds up. At the same time, the thickness d of the p-n junction is reduced because the respective space charges are formed in boundary layers of a progressively smaller thickness. Assuming an average impurity concentration in the case of germanium, we obtain $v_{\text{cont}} = 0.3$-0.4 V and $d = 10^{-4}$-10^{-5} cm. At the high impurity concentrations produced in some semiconductor devices, $v_{\text{cont}} \approx 0.7$ V and $d = 10^{-6}$ cm.

The diffusion of majority carriers across the junction is accompanied by the reverse migration of carriers under the influence of the electric field set up by the contact potential difference. This field moves holes from the n-region back into the p-region, and electrons from the p-region back to the n-region. In Fig. 2-1a, this drift of minority carriers is likewise shown by arrows. So long as the temperature of the specimen and of the surroundings remains unchanged, the p-n junction resides in a state of dynamic equilibrium. A definite number of electrons and holes diffuse every second across the boundary in the opposite directions, and as many of them are caused to drift by the field in the respectively reverse directions.

It is easy to think up a mechanical analogy of the process if we imagine that Fig. 2-1b shows a hill, and liken electrons to balls that go up this hill at various velocities corresponding to the thermal velocities of electrons. As they move uphill due to their initial velocities, the balls gradually slow down until they come to a complete stop at some definite height, and then roll downhill by gravity. This analogy holds for holes as well.

We have defined the motion of carriers due to diffusion as the diffusion current i_{dif}, and the motion of carriers due to the action of the field as the drift current i_{dr}. In a steady state, that is, in a state of dynamic equilibrium, the two currents are equal in magnitude and opposite in sign. Therefore, the net current across the p-n junction is zero, as it should be so long as no external voltage is applied to it. The two components may take on a range of values, depending on the carrier concentration and mobility. The height of the potential barrier is always such that a state of equilibrium is established. In other words, the diffusion current and the drift current completely balance out each other. This can be proved as follows. Suppose that for one reason or another (say, a rise in temperature), the rate of diffusion goes up. This will bring about an increase in the diffusion current, because a greater number of carriers will diffuse across the boundary. This will in turn build up the space charges and potentials on both sides of the boundary, the contact potential difference v_{cont} will rise, and this will mean that the electric field across the junction will gain in strength and a higher potential barrier will result. However, an increase in the field strength brings about a proportionate increase in the drift current, that is, in the reverse migration of carriers. So long as i_{dif} remains greater than i_{dr}, the potential barrier will go up in height. In the long run, however, the increase in i_{dr} will make it equal to i_{dif}, and v_{cont} will cease building up.

Figure 2-1c shows the distribution of carrier concentration in a p-n junction. This pattern is typical of germanium. Since the concentrations of majority and minority carriers differ by about six orders of magnitude or even more, the concentrations are laid off vertically on a logarithmic scale. The impurity concentrations in the n- and p-regions are ordinarily different. Exactly such a case is shown in Fig. 2-1c. It is taken that the majority and minority carrier concentrations respectively are $n_n = 10^{18}$ cm^{-3} and $p_n = 10^8$ cm^{-3} in the n-type material, while for the p-type material where the impurity concentration is lower the respective figures are $p_p = 10^{16}$ cm^{-3} and $n_p = 10^{10}$ cm^{-3}.

As is seen, the electron concentration in a p-n junction gradually varies from 10^{18} to 10^{10} cm^{-3}, and that of holes changes as gradually from 10^{16} to 10^8 cm^{-3}. As a result, what is known as the *depletion region* is formed at the interface between the two types of semiconductor. For example, at the very boundary, the electron concentration is 10^{14} cm^{-3} which is 1/10 000th of the figure in the *n*-region, and the hole concentration is 10^{12} cm^{-3} which is likewise 1/10 000th of the figure in the *p*-region. Obviously, the electric conductivity of the *p-n* junction must be a minute fraction of what it is in the remaining parts of the *n*- and *p*-regions.

The depletion region may alternatively be looked upon as an outcome of the action produced by the electric field which is set up by the contact potential difference. This field pushes mobile carriers from the boundary layers so that the electrons are moved into the *n*-region and the holes into the *p*-region.

This depletion region is also known as the *barrier region.* It has an extremely high resistance in comparison with the remaining portions of the *n*- and *p*-regions.

2-2 The Forward-Biased *P-N* Junction

Let an external voltage source be connected with its positive terminal to the *p*-type semiconductor and with its negative terminal to the *n*-type semiconductor that make up a *p-n* junction (Fig. 2-2*a*). This is called *forward biasing*, and the voltage producing it is referred to as the *forward bias voltage* v_f.

With forward biasing, the applied potentials establish an electric field which drives the majority carriers of each region towards the junction, thereby giving rise to the forward current, i_f, across the junction. This process is explained in Fig. 2-2*b*. (In this and the subsequent figures the potential diagram is shown simplified. Whatever happens in other parts of the circuit is of no interest as regards the *p-n* junction. Therefore, the diagrams do not show variations in potential along the *n*- and *p*-regions which implies that their resistance is arbitrarily assumed equal to zero. Nor do they show variations in potential at the contact of the *n*- and *p*-regions with the electrodes to which the wires from the voltage source are connected.)

The electric field set up in the *p-n* junction by the forward biasing voltage opposes the field due to the contact difference of potential. To reflect this fact, the vectors E_{cont} and E_f are shown pointing in the opposite directions. The resultant field is weakened and the potential difference at the junction is brought down which is another way of saying that the height of the potential barrier is reduced while the diffusion current builds up because a greater number of carriers can overcome the reduced barrier. At the same time, the drift current remains nearly unchanged because it mainly depends on how many minority carriers from the *n*- and *p*-regions are able to reach the *p-n* junction owing to their thermal velocities. If we neglect the voltage drop across the *n*- and *p*-regions, the voltage across the junction may be taken equal to $v_{\text{cont}} - v_f$. By way of comparison, the dashed line in Fig. 2-2*b* shows the potential diagram in the absence of an applied voltage.

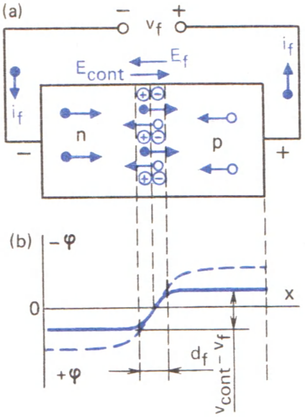


Fig. 2-2
Forward-biased *p-n* junction

As will be recalled, when no external voltage is applied to a *p-n* junction, i_{dif} and i_{dr} are equal in magnitude and opposite in sign so that they cancel each other. With forward biasing, $i_{\text{dif}} > i_{\text{dr}}$ and so the net current across the junction, that is, the forward current

$$i_f = i_{\text{dif}} - i_{\text{dr}} > 0 \tag{2-1}$$

If, on the other hand, the barrier is brought down appreciably, the diffusion current will be many times the drift current, and we may deem that the forward current is approximately equal to the diffusion current, which is another way of saying that the current across a forward-biased junction is a purely diffusion current.

The introduction of excess charge carriers across the potential barrier reduced in height by forward biasing into the region where they are minority carriers is called the *carrier* injection. The region of a semiconductor from which such excess minority carriers are injected is called the *emitter region* or, simply, the *emitter*. The region into which excess minority carriers are injected is called the *base region* or, simply, the *base*. If the injected minority carriers are electrons, then the n-region of a semiconductor will be the emitter and its p-region, the base. With the injection of holes, the situation will be the other way around: the p-region will be the emitter, and the n-region will be the base.

As a rule, the impurity concentrations and, in consequence, that of majority carriers in the n- and p-regions differ substantially. Therefore, the carrier injection is predominant from the region where these carriers are in majority. For example, if $n_n \gg p_p$, the injection of electrons from the n-region into the p-region will greatly exceed that of holes in the reverse direction. Accordingly, the n-region will act as an emitter, and the p-region as a base, because the injection of holes is negligible by comparison.

Forward biasing reduces both the height of the potential barrier and the thickness of the barrier layer (that is, $d_f < d$), so that its resistance in the forward direction, R_f, falls to a very small value (from units to a few tens of ohms).

With no external voltage applied, the barrier potential v_{cont} is just a few tenths of a volt. Therefore, the barrier height and the barrier resistance can be reduced to a very small value by applying as small a forward bias voltage (a few tenths of a volt). This is the reason why a heavy foward current can be produced with a very low forward bias voltage.

Obviously, there must be some forward bias voltage at which no potential barrier will be left in a p-n junction at all. Then the junction resistance, that is, the resistance of the barrier layer, will fall close to zero, and it may be neglected. In the circumstances, the forward current will be solely a function of the resistances of the n- and p-regions of the specimen. These resistances may no longer be neglected because they determine the magnitude of current across the junction. An example will illustrate this point best of all.

Suppose that in a forward-biased diode (for when an n-type and a p-type semiconductor are alloyed together, they form a single unit - a junction diode), the resistance of the barrier layer is 200 Ω, and the resistance of the n- and p-regions is 5 Ω each. Clearly, the total resistance of the diode is $200 + 2 \times 5 = 210$ Ω. This is very close to the resistance of the p-n junction alone, that is, 200 Ω. At some forward bias voltage the barrier disappears, the junction resistance falls to 0.5 Ω, and the total resistance ($0.5 + 2 \times 5 = 10.5$ Ω) may now be approximately considered to consist solely of two resistances of 5 Ω each. In other words, we may well neglect the resistance of the junction itself.

Now let us see how the forward current flows in the various parts of the circuit (Fig. 2-2*a*). Electrons leaving the n-region move across the junction into the p-region, and holes leaving the p-region move across the junction in the opposite direction into the n-region. Thus, we have two currents flowing at the same time: an electron current and a hole current. Of course, only the electron current can flow in the external leads of the circuit. They travel from the "−" side of the applied voltage source to the n-region and make up for the loss of electrons diffusing across the junction into the p-region. From the p-region, electrons travel towards the "+" side of the applied voltage source, leaving more holes behind in the process. This chain of events goes on non-stop, and so the forward current is flowing all the time.

The electron current is at its highest at the left-hand edge of the n-region. On moving closer to the transition region, this current falls off because an ever greater number of electrons recombine with the holes travelling across the junction towards the electrons, but the hole current, i_p, goes up. The total forward current, i_f, will be the same at any section of the circuit:

$$i_f = i_n + i_p = \text{const} \qquad (2\text{-}2)$$

This stems from the basic law for a series electric circuit. The same current is flowing in all parts of such a circuit.

Because the thickness of the barrier layer is very small and it is heavily depleted of charge carriers, very few of them recombine there, and the current in the barrier layer remains unchanged. Past the barrier layer, however, the electrons injected into the p-region recombine with holes. Therefore, as we move farther away from the junction to the right, that is, into the p-region, the electron current i_n keeps falling off, and the hole current i_p keeps building up. At the right-hand edge of the p-region, the electron current is a minimum, and the hole current is a maximum. Figure 2-3 shows how these currents vary along the x-axis for the case when the electron current exceeds the hole current because $n_n > p_p$ and the electron mobility exceeds the hole mobility. Of course, even in a forward-biased diode, the diffusion current is accompanied by the drift current produced by the motion of minority carriers, but it is negligibly small.

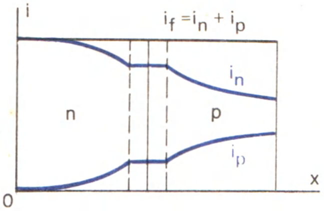


Fig. 2-3

Electron and hole current distribution in a *p-n* junction

2-3 The Reverse-Biased *P-N* Junction

Now we will connect the "+" side of an external voltage source to the *n*-region and the "−" side to the *p*-region of a *p-n* junction, or diode (Fig. 2-4*a*). Thus connected, the junction is said to be *reverse-biased.* The applied *reverse bias voltage* v_r causes a very small *reverse current,* i_r, to cross the junction. Why this current is small can be explained as follows. The field set up by the reverse bias voltage combines with the field due to the contact potential difference. In Fig. 2-4*a* this is indicated by the fact that the vectors $\vec{E}_{cont}$ and $\vec{E}_r$ point in the same direction. The resultant field gains in strength so that the height of the potential barrier now is $v_{cont} + v_r$ (Fig. 2-4*b*). Even a small rise in the barrier puts an end to the diffusion of majority carriers across the junction so that $i_{dif} = 0$ because the initial velocities of the carriers are not high enough for the carriers to overcome the barrier. In contrast, the conduction (drift) current remains nearly unchanged because it is mainly constituted by the minority carriers reaching the *p-n* junction from the *n*- and *p*-regions. The removal of minority carriers across a *p-n* junction by the accelerating electric field set up by the reverse bias voltage is called the *carrier extraction.*

Thus, the reverse current i_r is the conduction current produced by the movement of minority carriers. The reverse current is very small because minority carriers are small in number and, also, the resistance of the barrier layer in a reverse-biased *p-n* junction (or diode) is very high. The point is that a rise in the reverse bias voltage builds up the field at the junction, and a greater number of majority carriers are expelled by this field from the boundary layers on each side of the boundary into the *n*- and *p*-regions. Therefore an increase in the reverse bias voltage brings about an increase not only in the height of the potential barrier, but also in the thickness of the barrier layer ($d_r > d$). The barrier layer is depleted of carriers still more, and its resistance goes up appreciably so that $R_r \gg R_f$.

Even a relatively low reverse bias voltage causes the reverse current remain at some practically constant value. This is because the number of minority carriers is limited. A rise in temperature brings about an increase in their number so that the reverse current builds up but the reverse resistance goes down.

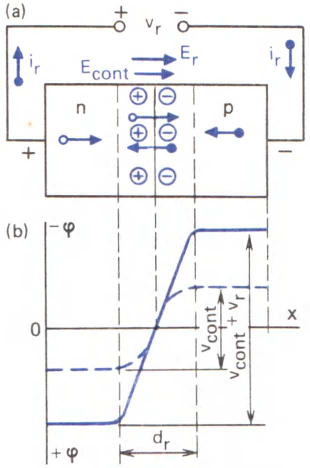


Fig. 2-4

Reverse-biased *p-n* junction

Let us take a closer look at how the reverse current behaves after a reverse bias voltage is applied. At first, this event gives rise to transients related to the movement of majority carriers. Electrons in the *n*-region move towards the "+" terminal of the applied voltage source, that is, away from the *p-n* junction. In the *p*-region, holes move away from the junction. At the "−" terminal, they recombine with the electrons still coming from the conductor connecting this electrode to the "−" terminal.

Because electrons leave the *n*-region, it is charged positively due to an excess of positively charged donor impurity atoms. The *p*-region is charged negatively because its holes are filled by the electrons coming there, and it acquires excess negatively charged acceptor impurity atoms.

The movement of majority carriers in the opposite directions we have just described goes on for a small span of time. This short-duration current is not unlike the charging current of a capacitor. Two unlike space charges are produced on either side of the *p-n* junction, and the entire system acts similarly to a charged capacitor with a poor dielectric and with a leakage current (its role in our case is played by the reverse current). In accord with Ohm's law, however, the leakage current of a capacitor is proportional to the applied voltage, but the reverse current across a *p-n* junction only slightly depends on the applied bias voltage.

2-4 The Metal-Semiconductor Junction

In addition to *p-n* junctions, present-day semiconductor devices use arrangements in which a metal and a semiconductor are brought into contact. Such an arrangement is called a *metal-semiconductor junction.* The events taking place in metal-semiconductor junctions depend on what is known as the *electronic work function* which is defined as the minimum energy required to liberate an electron from a metal or a semiconductor at absolute zero temperature. The lower the work function, the greater the number of electrons that can break loose from the specimen. Let us see how this happens in several types of metal-semiconductor junctions.

If in a junction formed by a metal and an *n*-type semiconductor (Fig. 2-5*a*) the electronic work function of the metal, Φ_M, is lower than the electronic work function of the semiconductor, Φ_S, escape of electrons from the metal into the semiconductor will be predominant. Therefore, electrons (majority carriers) are accumulated in the semiconductor layer next to the interface, and this layer becomes enriched with electrons – it has an increased concentration of electrons. The resistance of this layer will be low with either forward or reverse biasing, and so it is not capable of conducting current in one direction only. This is a *nonrectifying* (or *ohmic*) *contact.* A similar ohmic (or nonrectifying) contact exists when a junction is formed by a metal and a *p*-type semiconductor (Fig. 2-5*b*) if $\Phi_S < \Phi_M$. In this arrangement, more electrons will leave the semiconductor for the metal than the other way around. Likewise, a region enriched with holes (majority carriers) and having a low resistance is produced in the boundary region of the semiconductor. The two types of ohmic contact are widely used when attaching electrodes to the *n*- and *p*-regions of a semiconductor device.

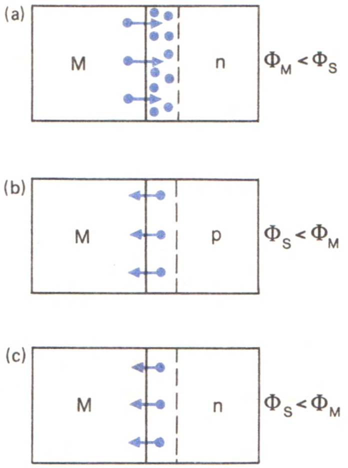


Fig. 2-5
Metal-semiconductor junction (*M* stands for 'metal')

The situation is different in the metal-semiconductor junction shown in Fig. 2-5*c*. Here, $\Phi_S < \Phi_M$. In the circumstances, electrons mainly move from the semiconductor into the metal, and the region produced near the interface on the semiconductor side is depleted of majority carriers and so it has a high resistance. As a result, a relatively high potential hill is formed here, with a height markedly varying according to the polarity of the applied voltage. This type of junction has the property of *unidirectional conduction* or *rectification.* This class of junctions was originally investigated by W. Schottky of Germany, and the potential hill appearing in such cases is called the *Schottky barrier*, and diodes utilizing it are called *Schottky diodes.* In the metal part of a Schottky diode, carrier storage (or the storage effect) is nonexistent in contrast to *p-n* junctions, and so Schottky diodes have a faster speed of response since there is no time lag associated with the storage time (defined as the time interval between the application of the reverse bias and cessation of the reverse current surge).

Similar rectifying properties are displayed when a metal is brought in contact with a *p*-type semiconductor, provided $\Phi_M < \Phi_S$.

Chapter Three

Semiconductor Diodes

3-1 The Current-Voltage Characteristic of the Semiconductor Diode

Whatever the type of electron device, it is important to know how its current varies with the applied voltage. Knowing this relation, we can readily find the current for any given voltage or to find the voltage for any given current.

If the resistance of a device is constant and independent of current or voltage, the two quantities may be connected by a relation of the form

$$i = v/R \quad \text{or} \quad i = Gv \tag{3-1}$$

As is seen, the current through a device is directly proportional to the applied voltage. The coefficient of proportionality is called the *conductance*

$$G = 1/R$$

A plot relating current to voltage is called the *current-voltage (volt-ampere)* characteristic, or simply, the characteristic of a device. For a device obeying Ohm's law, it is a straight line passing through the origin of coordinates (Fig. 3-1).

The higher the resistance R, the lower the conductance G and the smaller the current at the same given voltage. Therefore, at high values of resistance the characteristic makes a smaller angle with the axis of abscissas. The resistance R is connected to this angle α by a relation of the form

$$R = v/i = k \cot \alpha \tag{3-2}$$

where k is a proportionality coefficient which takes care of the units in which the quantities used in the equation are expressed, and the scale to which they are laid off along the coordinate axes.

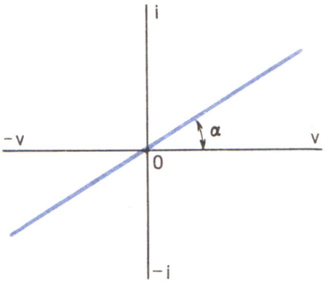


Fig. 3-1

Current-voltage characteristic of a linear device

Alternatively, we may write

$$G = 1/R = i/v = k' \tan \alpha \tag{3-3}$$

where $k' = 1/k$.

It is to be stressed that it would be wrong to write $R = \cot \alpha$ or $G = \tan \alpha$, because R and G are physical quantities that have particular dimensions and units with which they are expressed numerically, while $\tan \alpha$ and $\cot \alpha$ are trigonometric functions which are expressed only numerically. Also, the angle α may be different for the same value of R if we choose different scales on the axes. Devices obeying Ohm's law and having a straight-line current-voltage characteristic passing through the origin of coordinates are called *linear*.

There are devices whose resistance depends on voltage or current rather than remains unvarying. The relation between their current and voltage is not so straightforward as the simple Ohm's law, and their current-voltage characteristic is no longer a straight line passing through the origin of coordinates. Such devices are called *nonlinear*.

As already noted, when an *n*-type semiconductor is alloyed with a *p*-type semiconductor, they form a single unit called a *junction diode*. The nonlinear behaviour of such a diode is evident from reference to its current-voltage characteristic. As an example, Fig. 3-2 shows the current-voltage characteristic of a low-power diode. As is seen, a forward current of several tens of milliamperes is produced when the forward bias voltage is a few tenths of a volt. Therefore, the forward resistance is usually not greater than several tens of ohms. In the case of higher-power diodes, the forward current is hundreds of milliamperes or even greater at the same forward bias voltage, and R_f decreases to units of ohms or even less than that.

For the reverse current which is small in comparison with the forward current the characteristic is usually plotted on a different scale than for the forward current, as this is done in Fig. 3-2. For low-power diodes, a reverse bias voltage of several hundred volts gives rise to a reverse current of units or tens of microamperes. This corresponds to a resistance of several hundred kilohms or even greater. Since $v_r \gg v_f$, the two voltages are likewise laid off on different scales. Owing to this difference in scale, the curve has a kink or inflection at the origin of coordinates. If the same scale had been used, the curve would have had no kink.

The forward current characteristic shows an appreciable amount of nonlinearity because an increase in v_f leads to a decrease in the resistance of the barrier layer. Therefore, the curve runs with a progressively greater slope. At a voltage of a few tenths of a volt, the barrier layer practically disappears, and there remains only the resistance of the *n*- and *p*-regions, which may approximately be deemed constant. For this reason, the characteristic becomes almost linear. The small nonlinearity remaining here can be explained by the fact that the *n*- and *p*-regions are heated by the flow of current, and their resistance is thus brought down.

When the reverse voltage is raised, the reverse current first increases at a high rate. This happens because even at a low reverse voltage the rise in the height of the potential barrier brings about a sudden fall in the diffusion current which opposes the conduction (drift) current. In consequence, the total current, $i_r = = i_{dr} - i_{dif}$, abruptly goes up. Any further increase in the reverse voltage, however, entails only an insignificant rise in the reverse current. The current increases due to the heating of the junction by the current, due to the leakage over the surface, and also due to the avalanche multiplication of charge carriers, that is, the increase in the number of carriers as a result of impact ionization. Impact ionization consists in that at a high reverse voltage the electrons acquire a high velocity and, on striking the atoms of the crystal lattice, they knock out of it more electrons which are in turn accelerated and knock still more electrons out of the atoms. This process builds up with rising voltage.

At a certain value of reverse voltage, the *p-n* junction breaks down, leading to a sudden increase in the reverse current, and the resistance of the barrier layer abruptly decreases. A *p-n* junction may suffer two kinds of *breakdown, electric* and *thermal.* The electric breakdown whose region is labelled by the letters *ABC* in Fig. 3-2 is reversible. This means that it does not cause irreversible changes in the junction (the structure of the material remains unchanged). Therefore, semiconductor diodes may be operated under conditions of an electric breakdown. There are special diodes, sometimes called *breakdown diodes,* often used for voltage stabilization, which are designed to utilize the region *BC* of the characteristic.

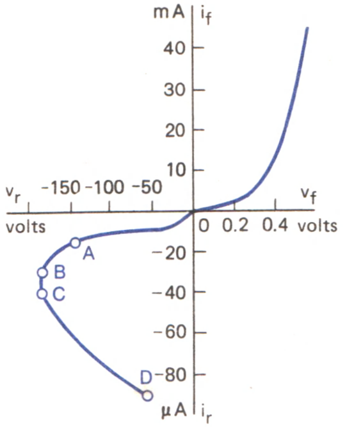


Fig. 3-2

Current-voltage characteristic of a semiconductor diode

Electric breakdown may in turn be classed into the avalanche type and the Zener type. *Avalanche breakdown* is caused by the cumulative multiplication of free charge carriers under the action of an applied field which brings about impact ionization and removal of electrons from the lattice atoms by the field. This type of breakdown is typical of thick *p-n* junctions produced at a relatively low impurity concentration in the host material. The breakdown voltage for avalanche breakdown is tens or hundreds of volts.

Zener breakdown is observed in a reverse-biased *p-n* junction that has a very high doping concentration on both sides of the interface. The built-in field is high (over 10^5 V cm^{-1}) and the depletion (or barrier) region is narrow as a result of the high doping level. The application of a small reverse voltage (just a few volts) is sufficient to cause electrons to tunnel directly from the valence band into the conduction band without changing their energy – this is the essence of what is known as the *tunnel effect*. In more detail, the tunnel effect is discussed in Chap. 8.

Thermal breakdown occurs within the region represented by portion *CD* of the curve in Fig. 3-2. This is an irreversible breakdown because it is accompanied by the destruction of the material at the *p-n* junction. Thermal breakdown occurs when the amount of heat dissipated at the junction due to the flow of reverse current exceeds the amount of heat withdrawn from the junction. As a result, the temperature of the junction rises, its resistance decreases, and the current through the junction builds up. The junction is thus overheated and destroyed by heat. An alternative name of this occurrence is, quite aptly, *thermal runaway*.

3-2 The Capacitance of a Semiconductor Diode

As is noted in Sec. 2-3, a reverse-biased *p-n* junction is not unlike a capacitor with a marked leakage current through the dielectric. The depletion layer has a high resistance and acts as a dielectric on each side of which there are two unlike space charges, $+Q_r$ and $-Q_r$, produced by the ionized atoms of the donor and acceptor impurities. Therefore, a *p-n* junction has a capacitance similar to that of a two-plate flat (plane-parallel) capacitor. It is called the *depletion-layer* or *barrier capacitance*. In the case of d. c. voltage, it is given by

$$C_b = Q_r/v_r \tag{3-4}$$

and in the case of a. c. voltage, by

$$C_b = \Delta Q_r/\Delta v_r \tag{3-5}$$

As with the capacitance of conventional capacitors, the barrier capacitance increases with increasing surface area of the *p-n* junction, increasing permittivity of the semiconductor and decreasing thickness of the depletion layer. Although in low-power *p-n* diodes the *p-n* junction has a small surface area, the barrier capacitance is fairly large because the depletion layer is narrow while the relative permittivity of the material is rather high (in the case of germanium, $\varepsilon = 16$). Depending on the surface area of the *p-n* junction, C_b may range from units to hundreds of picofarads. A distinction of the barrier capacitance is that it is nonlinear – it varies with changes in the voltage across the junction. When the reverse voltage is raised, the depletion layer broadens, and C_b decreases. The manner in which C_b varies as a function of v_r is shown by the plot in Fig. 3-3. As is seen, a change in v_r can bring about a three-fold change in C_b.

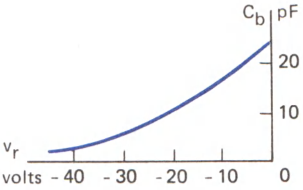


Fig. 3-3
Barrier capacitance as a function of reverse bias voltage

The barrier capacitance has a detrimental effect on the rectification of alternating current because it shunts the diode and thus provides a bypass for alternating current around it at high frequencies. However, the barrier capacitance can serve a useful purpose as well. For example, it is utilized in *varicaps* and *varactors* – *p-n* junction semiconductor diodes designed for low losses at high frequencies. They are used as tuning capacitors in tuned circuits and also in some other circuits which depend for their operation on the properties of nonlinear capacitance. In contrast to conventional variable capacitors in which the capacitance is changed mechanically, in varicaps this change is brought about by varying the reverse voltage applied. This control of tuned circuits is called *electronic tuning*.

When forward-biased, a *p-n* junction has what is known as the *diffusion capacitance*, C_{dif}, in addition to the barrier capacitance. It is likewise nonlinear and goes up with rising forward voltage, v_f. The diffusion capacitance is associated with the accumulation of mobile charge carriers in the *n*- and *p*-regions when the junction is forward-biased. So it practically exists only when a *p-n* junction diode is forward-biased, and a large number of carriers diffuse (are injected) over the reduced potential hill and, since they have no time to recombine, are stored in the *n*- and *p*-regions. If we assume that in a *p-n* junction diode the *p*-region acts as the emitter and the *n*-region as the base, then forward

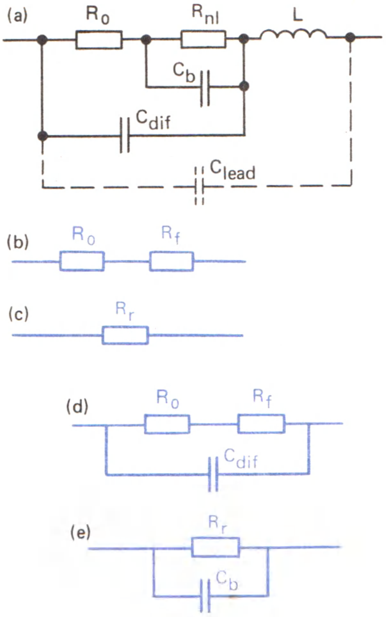


Fig. 3-4

Complete and simplified equivalent circuits of a semiconductor diode

biasing can cause a great number of holes* to move across the depletion layer from the p-region into the n-region so that a positive charge is produced in the n-region. At the same time, the d.c. voltage source causes electrons to move from the circuit conductor into the n-region and, as a consequence, a negative charge is formed there. The holes and electrons in the n-region cannot recombine instantaneously, and so for each value of forward voltage there is a certain value for the two equal but unlike space charges, $+Q_{\text{dif}}$ and $-Q_{\text{dif}}$, stored in the n-region owing to the diffusion of carriers across the junction. In the case of d.c. voltage, C_{dif} is the ratio of charge to potential difference:

$$C_{\text{dif}} = Q_{\text{dif}}/v_{\text{f}} \tag{3-6}$$

In the case of a.c. voltage, it is defined as

$$C_{\text{dif}} = \Delta Q_{\text{dif}}/\Delta v_{\text{f}} \tag{3-7}$$

As v_{f} is raised, the forward current builds up at a faster rate than the voltage because the current-voltage characteristic of a forward-biased p-n junction is nonlinear. Therefore, Q_{dif} rises faster than v_{f}, and C_{dif} increases.

The diffusion capacitance of a p-n junction is appreciably greater than its barrier capacitance, but there is no way of putting it to any use because it is shunted by the low forward resistance of the diode itself.

Recalling that a p-n junction diode has a capacitance, we may draw up its a.c. equivalent circuit as shown in Fig. 3-4 *a*. The resistance R_0 in this diagram is the total, relatively small resistance of the n- and p-regions, their respective electrodes and leads. When the diode is forward-biased, the nonlinear resistance R_{nl} is equal to R_{f}, that is, small. When the diode is reverse-biased, $R_{\text{nl}} = R_{\text{r}}$, that is, very high. This equivalent circuit may be simplified in many cases. At low frequencies the capacitive impedance is very high, and the diode capacitance may be neglected. Then the equivalent circuit will only include R_0 and R_{f} under forward bias (Fig. 3-4*b*), or only R_{r} under reverse bias (Fig. 3-4 *c*) because $R_0 \ll R_{\text{r}}$. At high frequencies, capacitances present a relatively low impedance. Therefore, the equivalent circuit under forward bias will take the form shown in Fig. 3-4 *d* (if the frequency is not very high, C_{dif} has practically no effect), while under reverse bias it also includes R_{r} and C_{b} (Fig. 3-4 *e*).

We should also include the capacitance between the diode leads, C_{lead}, which may markedly shunt the diode at very high frequencies. It is shown by the dashed lines in the figure. At microwave frequencies, the lead inductance may also have some effect.

3-3 The Temperature Behaviour of Semiconductor Diodes

Temperature has an appreciable influence on the electric conductivity of semiconductors. As the temperature goes up, a progressively greater number of electron-hole pairs is generated, the carrier concentration is raised, and so is the conductivity. This can be clearly seen from the current-voltage characteristics plotted at different temperatures. Figure 3-5 shows them for a germanium diode. As is seen, both the forward and reverse currents rise with increasing temperature. The increase is especially noticeable in the reverse current due to the increased genera-

* The flow of electrons from the n-region into the p-region may be neglected in this case because $n_n \ll p_p$.

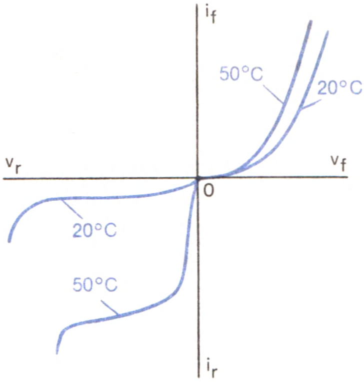


Fig. 3-5

Effect of temperature on the current-voltage characteristic of a semiconductor diode

tion of electron-hole pairs. In germanium diodes, the reverse current nearly doubles for every 10 degrees K rise in temperature. This can be written as

$$i_{r(t)} = i_{r(20^\circ C)} \times 2^{(t-20)/10} \tag{3-8}$$

Therefore, if the temperature goes up from 20°C to 70°C, the reverse current will increase 2^5 (that is, 32) times. Also, an increase in temperature brings down the breakdown voltage of germanium diodes.

In the case of silicon diodes, an increase of 10 degrees K in temperature brings up the reverse current by a factor of 2.5, while the breakdown voltage first increases somewhat with rising temperature, but then it falls off.

Heating does not raise the forward current of a diode in the same proportion as it does in the case of the reverse current. The explanation is that the forward current is mainly due to the extrinsic conduction, but the impurity concentration is temperature-independent.

A rise in temperature entails an increase in the barrier capacitance of the diode. The temperature coefficient of capacitance, TCC, which expresses the ratio of the change in capacitance to the original value for a unit change in temperature, is $10^{-3} = 10^{-4}\ \mathrm{K}^{-1}$.

3-4 The Operation of the Diode at Load

In practical circuits, a semiconductor diode always operates into some kind of load, say, a resistor (Fig. 3-6*a*). In circuit diagrams, the anode of a crystal diode is shown as a triangle, and its cathode as a bar. There is a flow of forward current when the anode is positive with respect to the cathode. Therefore, the triangle may be regarded as the arrowhead showing the conventional direction of forward current flow. It is the direction in which holes are moving under forward bias, while electrons are moving in the opposite direction.

The behaviour of a crystal diode in operation at load differs from that at no-load in several important respects. If the diode had a linear resistance, it would be a very simple matter to determine the current in such a case because the total resistance of the circuit is the sum of the d.c. resistance of the diode R_0 and of the load resistance R_L. Crystal diodes, however, are not linear resistances, and their R_0 varies with changes in the current that flows through. Therefore, the current around the circuit containing a crystal diode operating at load is found graphically. The problem may be stated thus: We know E, R_L and the diode characteristic and we are to find the current around the circuit and the voltage across the diode.

The diode characteristic should be looked upon as a plot of an equation connecting current i and voltage v. For R_L, a similar equation is Ohm's law:

$$i = v_R/R_L = (E - v)/R_L \tag{3-9}$$

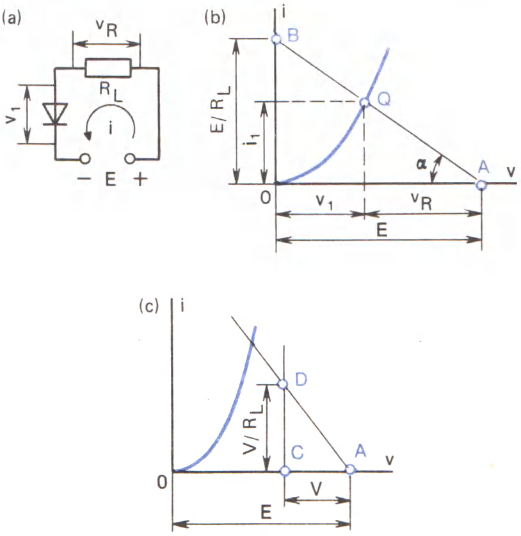


Fig. 3-6

Connection of a diode and load in circuit and construction of the load line

Thus, we have two equations in two unknowns, i and v, with one of the equations given graphically. To solve this set of equations, we need to construct a plot for the second equation and to locate the intersection of the two curves.

The equation defining R_L is a 1st-degree equation written in terms of i and v. Its plot is a straight line called the *load line*. The simplest way to construct it is to use two points on the coordinate axes. On setting $i = 0$, we get from Eq. (3-9):

$$E - v = 0 \quad \text{or} \quad v = E$$

which corresponds to point A in Fig. 3-6*b*. On setting $v = 0$, we get

$$i = E/R_L$$

We lay off this current as ordinate (point B), and join the two points, A and B, by a straight line. Thus, we get the load line. The coordinates of point Q yield the solution of our problem. It is to be noted that all the other points on line AB do not represent any operating conditions of the diode.

As an alternative, the load line can be constructed by using the slope angle α, because

$$R_L = k \cot \alpha$$

but this approach is less convenient because we would have to find the coefficient k so as to account for the scales used and to recover the angle α from its cotangent.

When the load line is constructed for relatively small values of R_L, point B may fall outside the drawing. If so, lay off an arbitrary voltage V to the left of point A (Fig. 3-6*c*) and, starting at point C thus obtained, lay off a current equal to V/R_L (segment CD). The straight line joining points A and D will then be the load line.

Sometimes, one may know v and i (point Q) and the load resistance R_L, and one is to determine E. Or one may know E and is to determine the load resistance R_L. We leave it as an exercise for the reader to construct the plots for these two cases. In either case, use Eq. (3-9) as the guide.

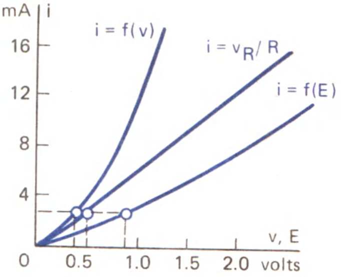


Fig. 3-7

Dynamic (or load) current-voltage characteristic for a circuit consisting of a series combination of a diode and a load resistor

A circuit containing a series-connected diode and a linear load resistance R_L is a nonlinear circuit. The characteristic of such a circuit, called the *dynamic characteristic of the diode*, that is, a plot of i as a function of E, $i = f(E)$, can be obtained by adding together the voltages taken from the characteristics of the diode and of the load resistance R_L (Fig. 3-7). The characteristic of the load resistance is in effect Ohm's law

$$i = v_R/R_L$$

and is a straight line passing through the origin of coordinates. It can be plotted as follows. Mark on the plot a point corresponding to an arbitrary value of v_R and v_R/R_L. Join this point to the origin of coordinates by a straight line. In the previous cases, the load line did not pass through the origin of coordinates because it related the current to the voltage v across the diode and not to the voltage v_R.

The dynamic characteristic of the circuit, $i = f(E)$, can be constructed by taking the sum of v and v_R for several values of current because $E = v + v_R$. As an example, at a current of 3 mA we have $v = 0.4$ V and $v_R = 0.5$ V. Adding the two voltages together, we obtain on the resultant curve a point corresponding to $E = 0.9$ V. Acting similarly, we locate other points and draw a smooth curve through them.

The behaviour of a series circuit mainly depends on the circuit section that has the highest resistance. Therefore, the greater the value of R_L, the less the nonlinearity of the resultant characteristic. It is to be noted that the operation of the diode at load need not be determined graphically if $R_L \gg R_0$. If so, it is legitimate to neglect the resistance of the diode and to determine the current approximately by the equation $i \approx E/R_L$.

The techniques we have used to find the d. c. voltage E may be used to determine both the peak and any instantaneous values when the anode source supplies an alternating voltage.

3-5 Semiconductor Diodes as Rectifiers

Rectification, or the conversion of alternating current into unidirectional or direct current, is one of the principal processes used in electronics.

Because semiconductor diodes have the property of unidirectional conduction (they conduct well in the forward direction and poorly, if at all, in the reverse direction), most of them are quite logically used for a. c. rectification.

An elementary rectifier circuit is shown in Fig. 3-8*a*. It is a series connection of a generator that supplies an a. c. emf, *e*, a diode *D*, and a load resistor R_L which may alternatively be placed in the other lead (as is shown by the dashed symbol in the diagram). The diagram shows what has come to be called a *half-wave rectifier* for the reason that it produces a pulsating current by passing only half the input cycle of an alternating current while the other half is blocked by the diode. It is a single-phase rectifier because the emf source is likewise single-phase. There are more elaborate rectifier circuits (two-phase, three-phase, etc.), but they are in effect a combination of several single-phase rectifiers.

In rectifiers used to energize electronic equipment, the a. c. source is usually a power transformer plugged into an a. c. power supply line (the a. c. mains) (Fig. 3-8*b*). Sometimes, autotransformers are used instead of transformers. In still other cases a rectifier may be connected to the mains directly, without any transformer. In practical applications, the role of the load resistor R_L shown in the diagram of Fig. 3-8 is played by the circuits or devices that are powered by the rectifier. In the case of r. f. rectification (or, more correctly, *r. f. detection*), such as in the detector stages of radio receivers, the a. c. emf source may be an r. f. transformer or a resonant tuned circuit, and the load is a high-value resistor.

An elementary rectifier operates as follows. Let us agree that the generator (source) supplies a sinusoidal emf

$$e = E_m \sin \omega t$$

and that its internal resistance may be neglected (if it may not be neglected, the internal resistance of the emf source should be accounted for in the usual way). During one half-cycle the diode is forward-biased, a current flows through it and produces a voltage drop v_R across the load resistor R_L. During the next half-cycle, the diode is reverse-biased, there is practically no current flowing, and $v_R = 0$. Thus, a pulsating current passes through the diode, the load resistor, and the source in the form of pulses each a half-cycle long and separated by intervals likewise a half-cycle long. This is the rectified current. It produces a rectified voltage across the load resistor R_L. If we trace the direction of current flow, we will readily establish its polarity or sign: the "+" terminal will be at the cathode and the "−" terminal at the anode.

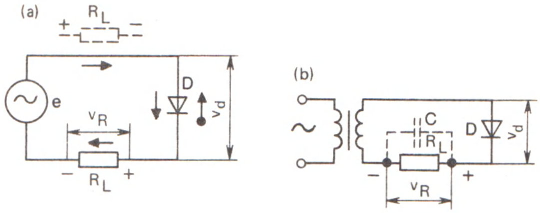


Fig. 3-8
Rectifier circuit using a crystal diode

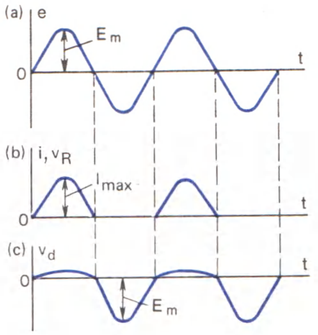


Fig. 3-9
Explaining the operation of a simple rectifier circuit

The plots in Fig. 3-9 give a clear idea about the events taking place in the rectifier. The a. c. emf supplied by the source is shown as a sinewave of amplitude E_m (Fig. 3-9*a*). As a rule, the load resistance is many times the diode resistance, and the nonlinearity of the diode may be neglected (the dynamic characteristic is close to linear). In the circumstances, the rectified current has the form of pulses, each pulse being nearly a half-sinewave of maximum value $I_{\max}$ (Fig. 3-9*b*). Drawn on another scale, the same plot will show the rectified voltage v_R because $v_R = iR_L$. Multiplying the value of current by R_L

will immediately yield a curve representing the rectified voltage.

The plot in Fig. 3-9*c* shows the voltage across the diode. Sometimes it is incorrectly regarded as sinusoidal or identified with the voltage supplied by the a. c. emf source. Actually, this is a nonsinusoidal voltage because its positive half-cycles markedly differ in amplitude from its negative half-cycles. The positive half-cycles have a very small amplitude. The explanation is that during the passage of forward current the greater proportion of source voltage is dropped across the load resistor whose resistance is many times the diode resistance. Therefore,

$$V_{\mathrm{f\,max}} = E_m - V_{R\,\max}$$
$$= E_m - I_{\max} R_{\mathrm{L}} \ll E_m \qquad (3\text{-}10)$$

For conventional semiconductor diodes, the forward voltage does not exceed 1 or 2 V.

Suppose that the source in our circuit supplies an rms voltage equal to $E = 200$ V and that $E_m = \sqrt{2}\,E = 280$ V. If $V_{\mathrm{f\,max}} = 2$ V, then $V_{R\,\max} = 278$ V. If the source voltage (say, 200 V) were fully impressed on the diode, this would mean that no voltage is dropped across R_{L}. This can only occur when $R_{\mathrm{L}} = 0$. Then the current would be excessively heavy, and the diode would be destroyed.

During the negative half-cycles of the applied voltage, there is practically no current flowing, and the voltage drop across R_{L} is zero very nearly. All of the source voltage is applied to the diode and biases it in the reverse direction. As a result, the maximum value of reverse voltage is equal to the amplitude of the source voltage.

Let us examine the rectified voltage in more detail (everything said about it fully applies to the rectified current). As is seen from the plot in Fig. 3-9*b*, the rectified voltage is a strongly pulsating one so that no voltage exists during one half of each cycle. The useful part of such a voltage is its *direct* (*constant*) or *average component*. It is often referred to simply as the average voltage and its symbol is V_{av}. For a half-sinewave pulse of maximum value $V_{\max}$, its average over a half-cycle is

$$V_{\mathrm{av}} = 2V_{\max}/\pi = 0.636 V_{\max} \qquad (3\text{-}11)$$

Since no voltage exists during the next half-cycle, the average over the entire cycle is half the previous value, or

$$V_{\mathrm{av}} = V_{\max}/\pi = 0.318 V_{\max} \qquad (3\text{-}12)$$

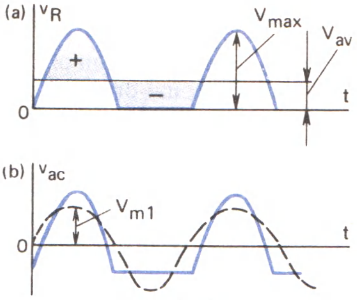


Fig. 3-10

Direct and alternating components of rectified voltage

Approximately, V_{av} is 30% of the maximum value. This approximation is quite legitimate because the actual pulses always differ in waveform from the half-sinewave. Because the voltage drop across the diode is very small, we may take it that

$$V_{\max} \approx E_m \text{ and } V_{\mathrm{av}} \approx 0.3E_m \qquad (3\text{-}13)$$

On subtracting the average value from the pulsating rectified voltage, we obtain its alternating component, V_{ac}, which is nonsinewave in shape. Its datum (or zero) axis is the straight line representing the direct (constant) component (Fig. 3-10*a*). The half-cycles of the alternating component are shown shaded. The positive half-cycle accounts for the upper two-thirds of the half-sinewave, and the negative half-cycle is close to a trapezoid in waveform. These half-cycles differ in duration, but they bound equal areas because they do not contain a direct component any longer.

The a. c. component is an 'unfavourable' part of the rectified voltage. Measures are usually taken to minimize it in the load resistor, that is, to smooth the pulsations, or *ripples*, in the rectified voltage. One way to do this is to use what are known as *smoothing filters*. They are also called *ripple filters* or *rectifier filters*. In Fig. 3-10*b*, the a. c. component is shown separately. It consists of several harmonics. The one most difficult to suppress is the first harmonic (or the fundamental), shown shaded in the figure.

A smoothing filter uses high-value capacitors which provide a bypass for the a. c. current so that as little of it could flow into the load as possible. As often, smoothing filters include chokes – high-value inductors which impede the flow of the a. c. component into the load. As the ripple (or pulsation) frequency goes up, the reactance presented by the filter capacitor decreases and that due to the filter inductors increases with the net result that the filter performance is improved.

If a filter performs well in suppressing the fundamental of the ripple, it will suppress the harmonics still better. Because the harmonics are smaller in amplitude than the fundamental, practically one only needs to take care of the fundamental, the worst "offender" of all.

In the elementary rectifier we have chosen as an example, the fundamental of the ripple is very large. Its amplitude V_{m1} is greater than that of the useful component:

$$V_{m1} = 0.5V_{\max} = 1.57V_{\mathrm{av}} \tag{3-14}$$

As a rule, a rectified voltage pulsating so strongly can hardly be put to practical use. More elaborate rectifier circuits do reduce the ripple somewhat. The simplest way to reduce the ripple, however, is to use a filter consisting of a high-value capacitor placed in shunt with the load resistor R_L (see Fig. 3-8*b*). The inclusion of a capacitor, however, affects the performance of the diode in a very substantial way.

A capacitor will smooth the ripple well if its capacitance is such that

$$1/\omega C \ll R_L \tag{3-15}$$

During some part of a positive half-cycle when the diode is forward-biased, there is a current flowing through the diode and charging the capacitor to a voltage close to E_m. During the time interval when no current is flowing through the diode, the capacitor discharges through the load resistor R_L and produces across it a voltage which falls off gradually. During every subsequent half-cycle the charge on the capacitor is restored, and the voltage across it rises again.

It takes very little time for the capacitor to charge via the relatively small resistance of the diode, but its discharge through the high-value load resistor is a far slower process. As a result, the voltage across the capacitor and across the load placed in shunt with it pulsates only slightly. Also, the capacitor builds up the direct component of the rectified voltage. In the absence of a capacitor, $V_{\mathrm{av}} \approx 0.3E_m$; with a capacitor of a sufficiently high value V_{av} comes very closely to E_m and may be equal to 0.8 or 0.95 of E_m or even greater. Thus, in a single-phase half-wave rectifier a capacitor increases the rectified voltage nearly three-fold. The greater the values of C and R_L, the slower the discharge of the capacitor, the smaller the ripple, and the closer V_{av} is to E_m. If we remove all of the load resistance from the circuit (operation at no-load, with $R_L = \infty$), a direct voltage free from any pulsations and equal to E_m will be developed across the capacitor.

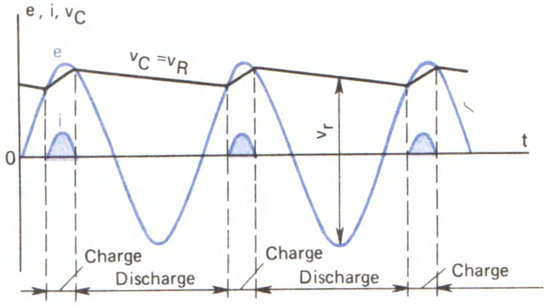


Fig. 3-11

Ripple reduction by a capacitor

The operation of a rectifier which contains a smoothing capacitor is explained in Fig. 3-11 showing the plots of the source emf *e*, the diode current *i*, and the capacitor voltage v_C equal to the load voltage v_R.

A better insight into what happens in a rectifier containing a capacitor may be provided by the following analogy. Suppose there is a machine which needs a steady and uniform supply of gas over a pipe. Unfortunately, the pump available to the operator can deliver the gas only portion-wise (similar to pulses in an electric circuit) because the pump draws in some gas only during the forward stroke of the piston and delivers it to the machine only during the reverse stroke. This system is not unlike a rectifier without a capacitor, with the pump motor acting similarly to the a. c. voltage source and with the pump valves acting as the rectifying diode. The situation can be improved if we install a large tank between the pump and the machine and fill it with gas. The tank will then supply the gas to the machine at a nearly constant pressure. It will pulsate only slightly because the pump will replenish the tank and maintain the average pressure in it at one and the same level. Thus, the tank operates similarly to a capacitor in a rectifier. The greater the size of the tank and the smaller the flow rate of gas to the machine, the smaller the pulsations in the gas pressure.

The " + " side of the capacitor is connected to the cathode and its " − " side to the anode of the diode. Therefore, the diode voltage v_{d} is equal to the difference between the source emf and the capacitor voltage

$$v_{\mathrm{d}} = e - v_C \tag{3-16}$$

Because v_C is nearly equal to E_m, the diode voltage becomes a direct one only during some part of a positive half-cycle when e exceeds v_C (near E_m). During these short intervals of time, the diode conducts a current in the form of pulses which restore the charge on the capacitor. During the remainder of each positive half-cycle and during the negative half-cycles, the diode voltage is reversed, there is no current flowing through the diode, and the capacitor discharges into R_{L}.

The reverse voltage across the diode is a maximum when the amplitude of the source emf is negative, $e = -E_m$. Because the voltage across the capacitor is then likewise close to E_m, the maximum reverse voltage is close in value to $2E_m$. When the load circuit is open-circuited (operation at no-load), the maximum reverse voltage is equal to $2E_m$ exactly. Thus, the use of a capacitor doubles the reverse voltage as compared with its value in the absence of a capacitor.* Therefore, it is important to choose a diode capable of standing up to this reverse voltage.

When the ripple must be kept to a very low minimum or when the load resistance is too small, a capacitor of a prohibitively high value would be required. In other words, the ripple could not then be smoothed by a capacitor alone, and one would have to add another smoothing filter consisting of a high-reactance choke and one more capacitor (or a still more elaborate filter).

It is essential to stress the danger associated with the short-circuit that may occur in the load when the filter capacitor is ruptured. Then all of the source voltage will be impressed on the diode, and it will be exposed to a prohibitively heavy current causing the thermal destruction of the diode.

Semiconductor diodes compare favourably with vacuum diodes not only because they need no filament (or heater) voltage for the cathode, but also because the voltage drop across the diode under forward bias is small. Whatever the current, or power, for which a semiconductor diode has been designed, its forward voltage is a few tenths of a volt or a bit higher than 1 V. Therefore, rectifiers using semiconductor diodes are more efficient than those using vacuum diodes. Importantly, the efficiency improves as the voltage to be rectified is increased because a loss of about 1 V across the diode itself is immaterial. For example, if the voltage to be rectified is 100 V and the voltage drop across the diode is 1 V, the efficiency is about 99% (the figure will of course be a bit smaller if we take into account some other losses).

Thus, semiconductor diodes are more economical than vacuum diodes and generate less heat in operation so that less damage is caused to the nearby components. Also, semiconductor diodes have a very long service life. They suffer from a disadvantage in that they can stand up to a relatively low reverse voltage – not more than several hundred volts while the figure for H. V. vacuum diodes may be tens of kilovolts.

Semiconductor diodes may be used in any type of rectifier circuit. If, however, the front-end element of a smoothing filter is a high-value capacitor, it might pass a current pulse which may exceed the limit of forward current for the diode. To avoid this, the usual practice is to connect the diode in series with a current-limiting resistor of several units or tens of ohms.

In diodes used as rectifiers, reversal of voltage polarity may give rise to substantial pulses of reverse current (Fig. 3-12). They can arise from two causes. Firstly, the reverse voltage gives rise to a current pulse which charges the depletion-layer (barrier) capacitance of the p-n junction – the higher this capacitance, the stronger the current pulse. Secondly, the reverse voltage permits the diffusion capacitance to discharge, that is, to release the minority carriers stored in the n- and p-regions. During the flow of forward current these carriers are injected across the interface and, failing to recombine or escape, are stored in the n- and p-regions. In practice, one has to reckon most of all with the large charge stored in the base region.

For example, if the electron concentration in the n-region is substantially greater than the hole concentration in the p-region, the n-region

* This is not the case with some rectifier circuits.

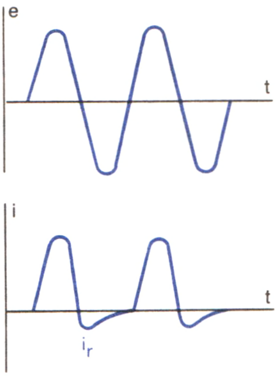


Fig. 3-12
Reverse current pulses in a crystal diode

will be the emitter and the *p*-region, the base. The injection of electrons from the *n*- into the *p*-region exceeds that of holes the other way around, so electrons are mostly stored in the *p*-region. When the voltage polarity is reversed, this charge is removed – the electrons move from the *p*- into the *n*-region, thereby giving rise to a reverse current pulse. The heavier the forward current, the greater the number of injected carriers (electrons in our case) and the greater the charge that they form. As a consequence, a stronger pulse of reverse current will be produced. After this charge has been removed and the barrier capacitance has practically been fully charged, there will remain an extremely small reverse current which may be neglected.

The reverse current pulse gains in strength with a rise in frequency. The explanation is that the reverse voltage builds up at a progressively faster rate as the frequency is raised. In consequence, the barrier capacitance is charged by a heavier current, that is, in a shorter span of time. In other words, a rise in frequency brings about a decrease in capacitive reactance, and the reverse current rises in proportion. The charge produced by the injected carriers is removed likewise faster, and this also serves to build up the reverse current pulse.

At low frequencies, the reverse current pulse is very small and its duration is only a small fraction of the half-cycle. At a certain high frequency the reverse current pulse may have about the same amplitude as the forward current pulse and last throughout the half-cycle. If the forward and reverse current pulses are equal in area, the direct component (the average rectified current) will be zero – there will be no rectification. In practice, it is recommended to use diodes for rectification up to a frequency at which the direct component of the rectified current drops by not more than 30% as compared with its value at low frequencies.

A rise in temperature brings about a decrease in R_f and R_r, but this usually has only a slight effect on rectification. The point is that the forward current is practically determined by the load resistance R_L which usually is many times the forward resistance of the diode, while the reverse resistance even of a hot diode remains sufficiently large in comparison with R_L, and so the reverse current remains small in comparison with the forward current.

The performance of diodes in low-frequency rectifiers can be evaluated in terms of several quantities. They include the forward current averaged over a period, $I_{f\,av}$; the corresponding voltage drop across the diode, $V_{f\,av}$, the reverse voltage, $V_{r\,av}$, and the corresponding reverse current, $I_{r\,av}$. The current $I_{f\,av}$ is often called the rectified current. Other important quantities are the maximum allowable (limiting) reverse voltage $V_{r\,max}$, the maximum allowable (limiting) forward (or rectified) current $I_{f\,max}$, the maximum safe case temperature $t_{case\,max}$, and also the operating frequency limit f_{max}.

3-6 Series and Parallel Connection of Diodes

When very high voltages are to be rectified, diodes have to be connected in series so that the reverse voltage across each diode could not exceed its safe limit. Unfortunately, it often happens that several diodes of the same type differ in reverse resistance (sometimes by a factor of tens). Because of this, the actual reverse voltage across some of the diodes may exceed their safe limit, and the diodes may break down. An example will give a better insight into the matter.

Suppose that in a rectifier the amplitude of reverse voltage is 1000 V and the diodes are chosen such that $V_{r\,max} = 400$ V. Obviously, at least three diodes have to be connected in series. Let the reverse resistances of the diodes be such that $R_{r1} = R_{r2} = 1$ MΩ and $R_{r3} = 3$ MΩ. The reverse voltage is divided in proportion to the reverse resistances, and so we find that $V_{r1} =$

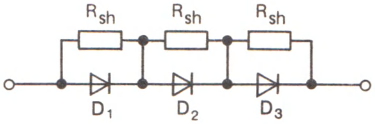


Fig. 3-13
Crystal diodes connected in series

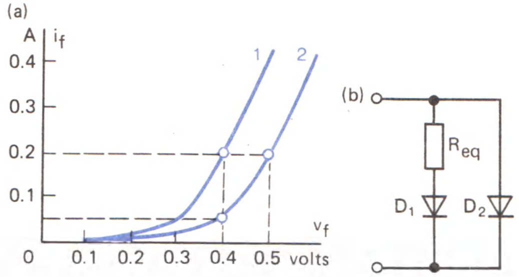


Fig. 3-14
Crystal diodes connected in parallel

$= V_{r2} = 200$ V and $V_{r3} = 600$ V. Thus, the reverse voltage across the third diode (which is, incidentally, the best of the three because it has the highest R_r) will exceed its safe limit, and the diode may break down. If this happens, the applied voltage (1000 V) will be divided among the remaining diodes, and a voltage of 500 V will exist across each of them. Obviously, any of the two may break down, and all of the 1000 V will then be applied to the remaining single diode, and the diode will not endure it. This chain of events sometimes happens in a matter of a split second.

For the reverse voltage to be divided equally among all the diodes irrespective of their reverse resistances, resort is made to resistors that are placed in shunt with the diodes (Fig. 3-13). The shunting resistors must have the same value of R_{sh} which is substantially smaller than the lowest of all the reverse resistances of the diodes. On the other hand, R_{sh} ought not to be too small, or else the reverse voltage will give rise to an excessive current, thus impairing the quality of rectification. For our example, we should take 100-kΩ resistors. Then, with a reverse voltage applied to the circuit, the resistance of each section of the circuit will be somewhat less than 100 kΩ, and the total reverse voltage will be divided among these sections in three nearly equal parts. The reverse voltage across each section will be lower than 400 V, and the diodes will operate reliably. As a rule, shunting resistors range in value from several tens to several hundred kilohms.

Diodes are connected in parallel when the desired forward current exceeds the current limit of a single diode. If, however, we connect several diodes of the same type simply in parallel, they will be loaded differently because of the spread in their volt-ampere characteristics, and the current in some of them will exceed the safe limit. The difference in forward current between diodes of the same type may be as great as tens of per cent.

As an example, Fig. 3-14*a* shows the characteristic of identical forward-biased diodes for which $I_{f\,max} = 0.2$ A. Suppose we want these diodes to deliver a direct current of 0.4 A. If we connect them in parallel, then at a current of 0.2 A the voltage across one diode will be 0.4 V (curve *1*), while the current in the other diode with the same voltage applied (curve *2*) will be a mere 0.05 A. Thus, the total current will be 0.25 A and not 0.4 A. The voltage across the diodes may not be raised because the current in the first diode would then exceed its safe limit.

It is seen from reference to the characteristics that if the second diode is to deliver a current of 0.2 A, the voltage across it must be 0.5 V which is by 0.1 V higher than it is across the first. Thus, for the diodes to operate properly, a voltage of 0.5 V must be applied and an equalizing resistor should precede the first diode (D_1) so as to absorb its excess 0.1 V at a current of 0.2 A (Fig. 3-14*b*). Obviously, the resistance of this resistor should be $R_{eq} = 0.1/0.2 = 0.5\ \Omega$. With this resistor connected as shown, the two diodes will be loaded identically by a current of 0.2 A.

In practice, it occurs but seldom that more than two or three diodes are connected in parallel. Equalizing resistors with a resistance of a few tenths of an ohm or units of ohms are usually chosen by trial and error until the same current is flowing in all the diodes in operation at load. Sometimes, the resistance of the equalizing resistors is deliberately chosen to be several times the forward resistance of the diodes. This is done in order that the current in each diode could be determined mainly by R_{eq}. Unfortunately, this brings about a further voltage drop across R_{eq}, which is many times the forward voltage across the diodes, and the efficiency is of course impaired. If it is not desirable to include equalizing resistors, the diodes to be used must be matched for their characteristics. Whenever possible, however, the parallel connection of diodes should preferably be avoided.

3-7 The Pulsed Operation of Semiconductor Diodes

Many state-of-the-art electronic circuits use semiconductor diodes in the pulsed mode with a pulse duration of several microseconds or even less. We will examine this mode of operation, taking as an example a diode connected in series with a load whose resistance, R_L, is many times the forward resistance of the diode, $R_L \gg R_f$.

Let such a circuit be acted upon by a pulsed voltage which consists of a short forward-voltage pulse (a positive-going pulse) which turns on the diode, and a longer reverse-voltage pulse (a negative-going pulse) which turns off the diode reliably until the arrival of the next positive (or enabling) pulse. The voltage pulses are rectangular in shape (Fig. 3-15*a*).

The waveform of current and of the proportional voltage across R_L is shown in Fig. 3-15*b*. When a forward voltage is impressed on the diode, the current in the circuit is determined by the load resistance R_L. Although the forward resistance of the diode is nonlinear, its effect is almost negligible because it is a small fraction of R_L. For this reason, the forward-current pulses will be left almost undistorted. A slight distortion may be observed only in the case of very short pulses (with a duration of a split microsecond).

Upon reversal of polarity, that is, when a reverse voltage is applied, the diode is not turned off at once, but during some time taken up by a reverse current pulse (Fig. 3-15*b*) which markedly exceeds in amplitude the reverse current in a steady state, $i_{r\,ss}$. This reverse-current pulse owes its origin to the same causes as when the diode operates at high frequencies (see Sec. 3-5). The primary cause is the discharge of the diffusion capacitance, that is, removal of the charges formed by mobile carriers in the *n*- and *p*-regions. Because the impurity concentrations in these regions are usually very different, the reverse-current pulse is mainly produced by removal of the charge stored in the base, that is, the region of a relatively low conductivity. For example, if the *n*-region acts as the emitter and the *p*-region as the base, the flow of holes in a forward-biased diode from the *p*- into the *n*-region may be neglected and only the flow of electrons from the *n*- into the *p*-region needs to be considered.

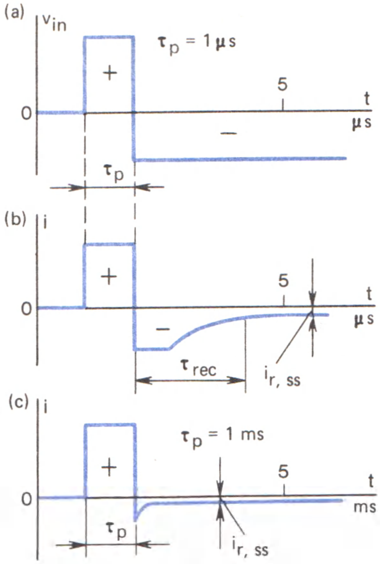


Fig. 3-15

Pulsed operation of a crystal diode

This diffusion current across the junction results in the accumulation and storage of electrons in the *p*-region because they cannot recombine or reach the *p*-terminal at once. Upon reversal of voltage polarity, the charge stored in the base region moves in the reverse direction, giving rise to a reverse current pulse. The heavier was the forward current, the greater the number of electrons stored in the base region and the stronger the reverse current pulse. On moving from the base back to the emitter, some of the electrons recombine with holes and some pass through the *n*-region and reach the metal electrode made to that region.

Removal of the space charge stored in the base lasts for some time. At the end of this time interval, the reverse current acquires its very small steady-state value, $i_{r\,ss}$. We may describe this chain of events somewhat differently. At first the reverse diode resistance R_r is relatively low, then it gradually rises and finally reaches its normal steady-state value. The time from the instant when a reverse current is produced to the instant when it falls to its steady-state value is called the *reverse recovery time*, τ_{rec}. It is a very important parameter for diodes intended for use in the pulsed (or switching) mode. For switching diodes the reverse recovery time does not exceed a split microsecond. The shorter this time, the better, because the diode will then take much less time to turn off.

The other cause for the generation of a reverse current pulse is the charging of the diode capacitance by the reverse voltage. The charging current of this capacitance is added to the current associated with removal of stored carriers and gives rise to a total reverse current pulse which increases in strength in proportion to the diode capacitance. For switching diodes, this capacitance does not exceed a few picofarads.

If the duration of the forward current pulse is substantially longer than that of the transients we have just examined (say, several milliseconds), the reverse current pulse will be extremely short and it may be neglected (Fig. 3-15*c*).

Apart from the reverse recovery time and diode capacitance, switching diodes are characterized by several other important quantities. They are the direct forward voltage V_f at a specified direct forward current I_f, the reverse current I_r at a specified reverse voltage V_r, the maximum allowable reverse voltage $V_{r\,max}$, and the maximum allowable forward current pulse $I_{f\,max}$.

3-8 Basic Types of Semiconductor Diodes

Semiconductor diodes may be classed into groups in many ways. One classification is by type of semiconductor material, another by frequency, still another by function or by construction, etc.

A very important classification is by type of structure. Here, all semiconductor diodes are classed into the *point-contact type* and the *junction type*. In point-contact diodes, the linear dimensions determining the area of the *p-n* junction are comparable with the thickness of the transition region or even smaller. In junction diodes, these dimensions are markedly greater than the thickness of the junction.

Point-contact diodes have a very small capacitance (usually less than 1 pF), and so they can be used at any frequencies, up to the microwave band. However, they can conduct a current of just a few units or tens of milliamperes. Junction diodes have a capacitance of tens of picofarads or even greater, depending on the junction area.

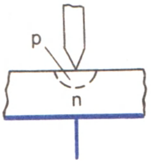


Fig. 3-16
Structure of a point-contact crystal diode

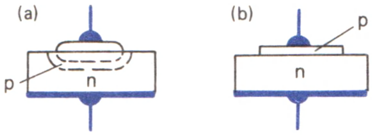


Fig. 3-17
Germanium *p-n* junction diodes manufactured by (*a*) alloying and (*b*) diffusion

The current limit for junction diodes may be as high as tens of milliamperes to hundreds of amperes or even greater.

Point-contact and junction diodes are fabricated from semiconductor wafers sliced off a single crystal which has a regular crystal structure in all directions. The source materials for point-contact and junction diodes are most often germanium and silicon. More recently, gallium arsenide (GaAs) and other compounds have come into use for this purpose.

In sketch form, the arrangement of a point-contact diode is shown in Fig. 3-16. It has a thin pointed wire (called a *catwhisker*) to which a desired impurity is applied and then is welded by a current pulse to a wafer of semiconductor having a certain type of conduction. When the metal point contacts the surface of the semiconductor, impurity atoms diffuse into the host material and produce a region of the opposite type of conduction (this step is called *forming*). In this way, a miniature hemispherical *p-n* junction is formed near the metal point. Hence the name "point contact".

Germanium point-contact diodes are fabricated from *n*-type germanium with a relatively high resistivity. The catwhisker welded to a germanium wafer is a tungsten wire given a coat of indium which acts as an acceptor impurity for germanium. The *p*-region thus produced in the germanium wafer acts as the emitter. Silicon point-contact diodes are fabricated from *n*-type silicon, and the catwhisker is given a coat of aluminium which behaves as an acceptor for silicon.

Junction diodes are mainly fabricated by *alloying* (or *fusion*) and *diffusion* (Fig. 3-17). The alloying process is a fabrication technique in which a small button of indium is fused at about 500°C into an *n*-type germanium wafer. The impurity metal alloys with the semiconductor material to form a *p*-type region of germanium. This *p*-type region has a higher impurity concentration than the remainder of the relatively high-resistance germanium, and so it acts as the emitter. Terminal leads, usually of nickel, are then welded to the germanium wafer and the indium button. If the host material is high-resistance *p*-type germanium, the impurity is usually antimony. On alloying with the host material, it forms an *n*-type emitter.

The alloying process produces what are called *abrupt p-n junctions* in which the region having a higher impurity concentration is substantially narrower than the region enclosing the space charges existing at the junction.

The diffusion process is a method of producing *p-n* junctions by disseminating acceptors or donors into a semiconductor at a high temperature. The impurity in the diffusion process is usually in the gaseous state. For the diffusion to proceed at a high rate, the host semiconductor is heated to a far higher temperature than in the alloying process. For example, an *n*-type germanium wafer will be heated to 900°C and placed in an atmosphere of gaseous indium. This treatment produces a layer of *p*-type germanium on the surface of the wafer. By varying the diffusion time, one can readily produce any desired thickness of such a layer with sufficient accuracy. On cooling, it is etched away from all parts of the wafer except one face. The diffused layer acts as the emitter. A diffused-junction diode is likewise fitted with electrodes and terminal leads which are made to the diffused layer and to the host wafer. In diffused-junction diodes, the impurity atoms penetrate to a relatively large depth into the host material, and so what is called a *graded p-n junction* is produced. In such diodes, the region where the impurity concentration varies is comparable in thickness with the region enclosing the space charges concentrated at the junction.

Now we will take a look at the various semiconductor diodes from the view-point of the functions they are intended to perform.

Rectifying junction diodes. Wide use is made of low-frequency rectifying junction diodes intended to rectify alternating current at frequencies to several kilohertz (sometimes, as high as 50 kHz). These diodes are used in rectifiers which power various equipment. Sometimes they are called power diodes. Low-frequency junction diodes are made of germanium or silicon. They are classed into low-power, medium-power, and high-power. The respective limits of rectified currents are 300 mA, from 300 mA to 10 A, and over 10 A. The diode ratings are usually quoted for operation at an ambient temperature of 20 ± 5°C.

Germanium diodes are usually fabricated by fusing indium into *n*-type germanium. They can stand up to a current density of as high as 100 A cm^{-2} at a forward voltage of up to 0.8 V. The limit of reverse voltage for them does not exceed 400 V, and the reverse current usually is a few tenths or hundredths of a milliampere for low-power diodes and several amperes for medium-power diodes. The operating temperature for these diodes ranges from − 60° to + 75°C. If a diode has to be operated at an ambient temperature in excess of 20°C, its reverse voltage must be brought down. The devices are likely to be overheated at a reduced atmospheric pressure or in the case of poor cooling. In the circumstances, overheating can be avoided by running the devices at a reduced rectified current.

High-power germanium diodes use natural air cooling. They are designed for a rectified current of as high as 1000 A and a reverse voltage of up to 150 V.

Of late, there has been a steadily growing interest in rectifying silicon diodes. They are fabricated by fusing aluminium into *n*-type silicon or an alloy of tin and phosphorus or gold into *p*-type silicon. They can also be fabricated by the diffusion process. Silicon diodes offer a number of advantages over germanium diodes. Their limit of forward current density may be as high as 200 A cm^{-2}, and the limit of reverse voltage, up to 1000 V. The operating temperature ranges from − 60° to + 125°C (or even + 150°C for some makes). The forward voltage of silicon diodes may run as high as 1-1.5 V which is somewhat greater than the figure for germanium diodes. The reverse current of silicon diodes is substantially lower than it is in germanium diodes.

High-voltage rectifiers use silicon piles enclosed in rectangular plastic cases which are in turn sealed with insulating resin. They can be designed for a current of several hundred milliamperes and a reverse voltage of several kilovolts. Silicon piles can be put together into larger units which can readily be assembled into various rectifier ciruits (such as bridge rectifiers or voltage doublers). Each pile in such a unit has terminal leads of its own for convenience in voltage adjustment. There are high-power silicon diodes which can handle rectified currents from 10 to 500 A at a reverse voltage of 50 to 1000 V.

Rectifying point-contact diodes. They are widely used at high frequencies and some of them even at microwave frequencies (up to several hundred megahertz) and can perform well at low frequencies. These diodes are extremely versatile as they may be used in a large variety of circuit configurations. Germanium and silicon point-contact rectifying diodes may be designed for a maximum allowable reverse voltage of up to 150 V and a maximum allowable rectified current of up to 100 mA.

Switching diodes. The manner in which these diodes operate and the quantities associated with this mode of operation have been discussed in Sec. 3-7. The most important quantity which decides whether a given diode can be used in applications involving short pulses is the reverse recovery time, τ_{rec}. In order to make it as short as possible, switching diodes are fabricated so that the junction capacitance is small and the carriers can recombine fast. Switching diodes are made for pulse currents up to several hundred milliamperes and a maximum allowable reverse voltage of a few tens of volts.

Applications involving very short pulses use what are known as *mesa diodes* (from the Spanish 'mesa' for 'table' or 'plateau'). They are fabricated by a process which turns them out in large numbers at a time. As the first step in the process, a wafer of the host semiconductor is given a layer of the opposite type of conduction by the diffusion process. During the second step, a mask is deposited over the diffused layer so as to protect a multiplicity of small areas against the subsequent etching. The unmasked areas are then etched away, and the masked areas appear as plateaus above the remaining material (Fig. 3-18) to act as small *p-n* junctions. Finally, the wafer is sliced into chips each of which is a junction diode. A distinction of mesa diodes is a reduction in the volume of the base region. As a result, the storage time of a switching mesa diode is markedly cut down. Since several diodes are made from a single wafer, the spread in characteristics and parameters between them is likewise minimized.

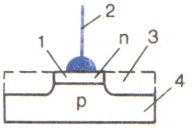


Fig. 3-18

Mesa diode: (*1*) diffused *n*-type layer; (*2*) *n*-region lead; (*3*) material removed by etching; (*4*) *p*-type semiconductor substrate

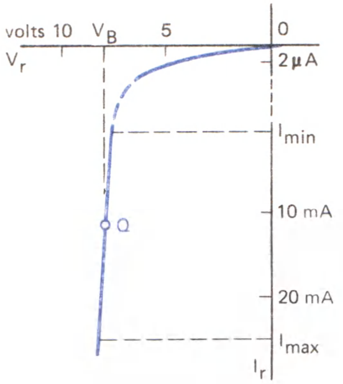


Fig. 3-19

Reverse current-voltage characteristic of a silicon breakdown diode

Breakdown diodes. As has been shown, the current-voltage characteristic of breakdown diodes has a portion which may be used for voltage regulation (or stabilization). In silicon junction diodes, this portion corresponds to variations in the reverse current between broad limits. Before a breakdown occurs, the reverse current is very small, at breakdown (in the voltage stabilization mode) it is comparable in magnitude with the forward current. As of this writing, many types of breakdown (voltage reference, voltage regulator, or VR) diodes are available, but they all are made of only silicon. They owe the name 'voltage reference' or 'voltage regulator' to the fact that once the breakdown has occurred, they will maintain their output voltage at a constant value that can be utilized as reference. Figure 3-19 shows the current-voltage characteristic of a typical breakdown diode under reverse bias. As is seen, when the diode is operating in the avalanche breakdown region (the voltage regulation mode), the output voltage changes very little. Under forward bias, the current voltage characteristic is the same as it is for conventional diodes.

Silicon VR diodes may be made for very low voltages (a few volts) such as are used to power many transistor circuits.

The basic parameters of silicon breakdown diodes include the following quantities. The breakdown voltage V_B may range from about 5 to 200 V, the diode current may range from tens to hundreds of milliamperes. The peak power dissipation P_{max} is from hundreds of milliwatts to several watts. The dynamic resistance $R_{dyn} = \Delta v/\Delta i$ in the VR mode may range from a few tenths of an ohm for low-voltage high-power breakdown diodes to 100-200 Ω for high-voltage devices. The dynamic resistance of low-voltage, low-power breakdown diodes is from a few ohms to tens of ohms. The lower the dynamic resistance, the better the performance of the diode. In an ideal case, $R_{dyn} = 0$. Since R_{dyn} is an a. c. resistance, it ought not to be confused with the d. c. resistance of a breakdown diode, defined as $R_0 = v/i$. The value of R_0 is always many times that of R_{dyn}. The temperature effect is stated in terms of the temperature coefficient of breakdown voltage, TCV, which is the ratio of the change in V_B to a unit change in temperature

$$\mathrm{TCV} = \Delta V_B/(V_B \Delta T) \tag{3-17}$$

The temperature coefficient of voltage may range from 10^{-5} to 10^{-3} K^{-1}. The value of V_B and the sign of the TCV depend on the resistivity of the host semiconductor. Breakdown diodes for voltages up to 6-7 V are made of silicon of a low resistivity, that is, with a high impurity concentration. In such diodes, the *p-n* junction is very narrow, the inherent field is very strong, and the breakdown mainly occurs by the Zener mechanism. This results in a negative temperature coefficient of voltage. When the source material is silicon with a lower impurity concentration, a wider *p-n* junction will be produced. It will break down at a higher voltage by the reverse-bias avalanche mechanism. Such diodes have a positive temperature coefficient of breakdown voltage.

Figure 3-20 shows a simple circuit using a breakdown diode. The load is placed in shunt with the diode. Therefore, in the breakdown region when the voltage across the diode remains nearly constant, the same voltage will exist across the load. If the source of voltage E is unstable, any changes in E will almost completely be absorbed by a limiting resistor, R_{lim}.

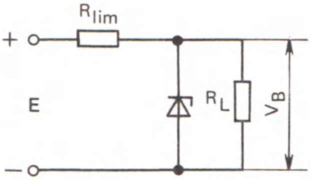


Fig. 3-20

Connection of a breakdown diode in a circuit

Most often, breakdown diodes are used in situations where the source voltage is unstable and the load voltage must be constant. To achieve proper voltage regulation, a certain definite value must be chosen for R_{lim}. As a rule, R_{lim} is calculated for the Q-point lying midway on the diode characteristic. If the source voltage E varies from E_{min} to E_{max}, the value of R_{lim} may be found by the following equation;

$$R_{lim} = (E_{av} - V_B)/(I_{av} + I_L) \tag{3-18}$$

where $E_{av} = (E_{min} + E_{max})/2$ is the arithmetic mean of source voltage, $I_{av} = (I_{min} + I_{max})/2$ is the arithmetic mean of diode current, and $I_L = V_B/R_L$ is the load current.

Should E change one way or the other, the diode current would also change, but the voltage across it and, in consequence, the voltage applied to the load would remain constant very nearly.

Because any changes in the source voltage must be absorbed by the limiting resistor, the maximum change in source voltage equal to $E_{max} - E_{min}$ must correspond to the maximum possible change in current at which the breakdown diode still retains its voltage regulation ability, that is, $I_{max} - I_{min}$. It follows then that if E changes by ΔE, voltage regulation will be effected on satisfying the condition

$$\Delta E \leqslant (I_{max} - I_{min}) R_{lim} \tag{3-19}$$

Voltage regulation in the case of larger changes in E can be retained by increasing the value of R_{lim}. It follows from Eq. (3-18) that a higher value of R_{lim} entails a lower value of I_L, that is, a higher value of $\mathrm{R_L}$. An increase in E_{av} likewise leads to an increase in R_{lim}.

Sometimes it may be necessary to obtain

a regulated voltage lower in value than that supplied by the breakdown diode used. This goal can be achieved by placing the load in series with a small resistor whose resistance can readily be found by Ohm's law (Fig. 3-21).

Another case of voltage regulation occurs when E is constant and R_L ranges from $R_{L\,min}$ to $R_{L\,max}$. For this case, R_{lim} can be found from the average values of currents, using the following equation:

$$R_{lim} = (E - V_B)/(I_{av} + I_{L\,av}) \qquad (3\text{-}20)$$

where

$$I_{L\,av} = (I_{L\,min} + I_{L\,max})/2$$
$$I_{L\,min} = V_B/R_{L\,max}$$
$$I_{L\,max} = V_B/R_{L\,min}$$

The operation of the circuit in the above case may be explained as follows. Because R_{lim} is constant and the voltage drop across it, equal to $E - V_B$, is likewise constant, the current in R_{lim}, equal to $I_{av} + I_{L\,av}$, must also be constant. But this is possible only if the diode current I and the load current I_L change to the same extent but in opposite senses. For example, if I_L rises, the diode current I must fall by the same amount so that their sum remains unchanged.

Where high regulated voltages are needed or involved, resort is made to a series connection of breakdown diodes designed each for the same current (Fig. 3-22). It is not recommended to use parallel connection of several breakdown diodes in order to obtain a higher regulated voltage because individual devices even of the same type may greatly differ in characteristics and parameters. Parallel connection may be used only if the total power dissipation by all the diodes does not exceed the peak power of a single diode.

As an alternative, breakdown diodes may be connected in cascade (Fig. 3-23) in which case diode D_1 must have a higher V_B than diode D_2.

How well voltage regulation is performed is stated in terms of the voltage regulation ratio (or the stabilization factor), k_{reg}. It is defined as the ratio of a fractional change in source voltage to the fractional change in the breakdown voltage. For the simple circuit shown in Fig. 3-20 we may write

$$k_{reg} = (\Delta E/E)/(\Delta V_B/V_B) \qquad (3\text{-}21)$$

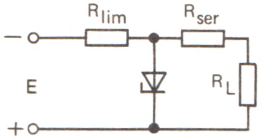


Fig. 3-21

Connection of a series resistor to bring down the regulated voltage across load

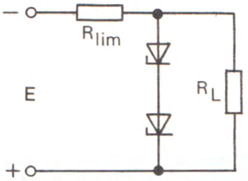


Fig. 3-22

Breakdown diodes connected in series

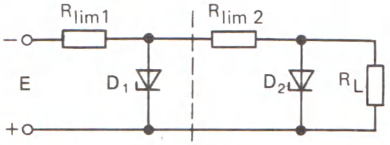


Fig. 3-23

Cascade connection of breakdown diodes

Practical breakdown diodes have a k_{reg} of several tens. For a cascade connection the overall k_{reg} is the product of the individual k_{reg}:

$$k_{reg\,tot} = k_{reg\,1}\,k_{reg\,2} \dots \qquad (3\text{-}22)$$

and may be as high as several hundreds even with two stages in cascade.

The voltage regulation schemes examined above suffer from a drawback which consists in that a good deal of power is dissipated in the diode itself and in the limiting resistor(s), with the result that the efficiency is heavily impaired. The losses are especially noticeable in a cascade connection.

It is to be noted that if the source voltage E is subject to fluctuations or pulsations, a breakdown diode will smooth them out to a great extent. This is because a breakdown diode has a low a. c. resistance which is usually a small fraction of R_{lim}. Therefore, the greater proportion of the ripple voltage will be absorbed in R_{lim}, and only a very small fraction of this

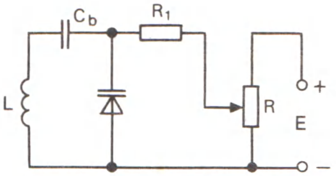


Fig. 3-24

Connection of a varactor in a resonant circuit as a variable capacitor

voltage will be dropped across the breakdown diode and the load.

Varactors. A varactor is a *p-n* junction semiconductor diode which utilizes variations in the junction capacitance with reverse bias. In effect, varactors are variable capacitors whose capacitance is controlled electrically (by varying the reverse voltage) rather than mechanically.

Varactors are mainly used as tuning elements in resonant (tuned) circuits and some specialized devices, such as parametric amplifiers. Figure 3-24 shows a diagram of a simple tuned circuit containing a varactor. By varying the reverse voltage across the varactor with a potentiometer, *R*, we can control the resonance frequency of the tuned circuit. R_1 is a high-value series resistor included in order to maintain the *Q*-factor of the tuned circuit in the face of the shunting effect of the potentiometer *R*. C_b is a d.c. blocking capacitor; without it the varactor would have been short-circuited for direct current by the tuned-circuit inductor *L*.

The job of varactors can well be done by silicon breakdown diodes operated at a voltage below V_B when the reverse current is still very small and, in consequence, the reverse resistance is very high.

We have examined only the most commonly used types of semiconductor diodes. There are also a number of special-purpose diodes some of which will be described in Chap. 8.

Chapter Four
Bipolar Transistors

4-1 General Principles

Transistors are semiconductor devices capable of power amplification and having three or more leads. They may have two or more *p-n* junctions, but most common among them are those with two *p-n* junctions. They are called *bipolar junction transistors*. They owe the name ‘bipolar’ to the fact that they utilize the flow of both minority and majority charge carriers through the device.

The first transistor was invented in 1948. It was a point-contact transistor. The point-contact transistor consisted of a small crystal of semiconductor (usually germanium) with two rectifying point contacts attached in close proximity to each other and a single large-area ohmic contact at some distance from the point contacts. Unfortunately, point-contact transistors have proved unstable in operation and are now obsolete, being ousted by junction transistors which were invented in 1949.

The basic principles of a bipolar junction transistor are illustrated in Fig. 4-1. It is a wafer of germanium, silicon, or some other semiconductor in which three regions differing in the type of conduction have been produced. As an example, we have taken an *n-p-n* transistor in which the middle region has hole conduction, and the two outer regions have electron conduction. As common are *p-n-p* transistors in which the outer regions have hole conduction and the middle region, electron conduction.

The middle region is called the *base*, one of the outer regions is called the *emitter*, and the other, the *collector*. Thus, our transistor has two *p-n* junctions, one between emitter and base, and the other between base and collector. The spacing between them must be very small, not more than a few micrometers and this means that the base region must be very narrow. This requirement is essential for proper operation of a transistor. Also, the impurity concentration in the base is always substantially lower than it is in the

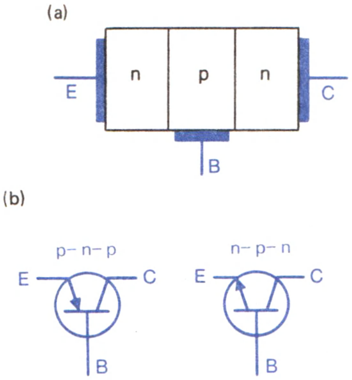


Fig. 4-1
Structure and graphical symbol of a *p-n* junction transistor

collector and emitter regions. Attached to each of the three regions is a metal electrode and a lead.

The quantities associated with the base, emitter, and collector carry, respectively the subscript B, E, or C. For example the base, emitter and collector currents are designated as i_B, i_E and i_C. The voltages between the electrodes are usually supplied with two-letter subscripts. For example, the voltage between base and emitter is labelled as v_{BE}, and that between collector and base as v_{CB}. In the diagram (or circuit) symbols of *p-n-p* and *n-p-n* transistors the arrow shows the conventional direction of current flow (from the " + " to the " − " terminal) in the emitter lead when the emitter-base junction is forward-biased.

A transistor can operate in any one of three regions, depending on the voltages across its junctions. These are the *active region*, the *cutoff region*, and the *saturation region*. A transistor is said to be operating in the active region when its emitter-base junction is forward-biased and its collector-base junction is reverse-biased. A transistor is driven into the cutoff region by reverse-biasing both *p-n* junctions. When, on the other hand, the two junctions of a transistor are forward-biased, it is said to be operating in the saturation region or *at saturation*. The active region is the basic mode of operation. It is utilized in most amplifiers and oscillators. Therefore, we will discuss the operation of the transistor in the active region in more detail. The cutoff and saturation regions are typical of transistors operating in the switching mode, and they will also be discussed, but at a later time.

As a rule, in any application using a transistor, two circuits are formed. One is the *input* or *control circuit* and the other is the *output* or *controlled circuit*. The source of the signal to be amplified is connected into the input circuit, and the load is connected to the output circuit. The quantities associated with the input circuit may carry either the letter subscript "in" or the numerical subscript "1". The quantities associated with the output circuit may be labelled "out" or "2".

4-2 Physical Processes in a Transistor

To begin with, we will see how, say, an *n-p-n* transistor operates with its load disconnected (static operation or operation at no-load), when only sources of direct supply voltages, E_1 and E_2, are connected into the circuit (Fig. 4-2*a*). Their polarity is such that the emitter junction is forward-biased, and the collector junction is reverse-biased. Therefore, the resistance of the emitter junction is low, and for a normal current to flow across this junction it will suffice to apply a voltage, E_1, of a few tenths of a volt. The resistance of the collector junction is high, and E_2 is usually from several volts to tens of volts. As is seen from the circuit diagram of Fig. 4-2*a*, the transistor voltages are connected by a simple relation of the form

$$v_{CE} = v_{CB} + v_{BE} \qquad (4\text{-}1)$$

When a transistor is operating in the active region, it is usually always that v_{BE} is substantially lower than v_{CB}, and in consequence v_{CE} is approximately equal to v_{CB}.

The current-voltage characteristic of the emitter junction is in effect the characteristic of a forward-biased semiconductor diode (see Fig. 3-2) and the current-voltage characteristic of the collector junction is similar to that of the reverse-biased semiconductor diode.

In basic terms, the operation of a transistor consists in that the forward voltage across the emitter junction, v_{BE}, has a strong effect on the collector current: the higher this voltage, the heavier the emitter and collector currents, and variations in the collector current are only slightly lower than those in the emitter current. In this way, the emitter-base voltage, v_{BE}, which is the input voltage, controls the collector

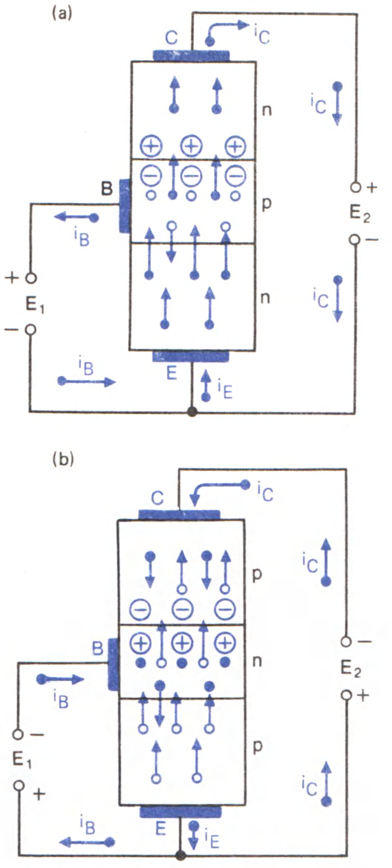


Fig. 4-2

Motion of electrons and holes in an *n-p-n* and a *p-n-p* transistor

current. This is the basis of signal amplification by transistors.

Physical processes occur in a transistor as follows. An increase in the forward input voltage v_{BE} brings about a fall in the height of the potential barrier at the emitter junction and a proportionate increase in the current flowing across that junction, that is, in the emitter current i_E. The electrons that make up this current are injected from the emitter into the base and also diffuse through the base into the collector region, thereby boosting the collector current. Because the collector junction is reverse-biased, there appear space charges shown as encircled " + " and " − " signs in the figure, and an electric field is set up between them. This field assists in sweeping the electrons arriving here from the emitter across the collector junction or pulling them into the collector region.

If the base is narrow enough and its hole concentration is low, most of the electrons swept across the base will not have time to recombine with holes in the base, but will reach the collector junction. Very few of the electrons recombine with holes in the base. Recombination gives rise to the base current that flows in the base lead. In a steady state, the number of holes in the base must remain unchanged. Because of recombination, so-many holes disappear every second, but as many new holes are produced owing to the fact that an equal number of electrons leave the base for the " + " terminal of the E_1 source. In other words, the base cannot retain very many electrons. If some of the electrons injected into the base from the emitter fail to reach the collector and remain in the base to recombine with holes, exactly as many electrons should leave the base as the base current i_B. Because the collector current is smaller than the emitter current, the following relation always exists between the currents in accord with Kirchhoff's current (or first) law:

$$i_E = i_C + i_B \qquad (4\text{-}2)$$

The base current is useless if not detrimental. Preferably, it should be as small as practicable. As a rule, i_B accounts for a few per cent of the emitter current, $i_B \ll i_E$, and so the collector current is only slightly lower than the emitter current. In other words, we may write that i_C is approximately equal to i_E. It is for the purpose of making the base current as small as possible that the base region is made very narrow and its impurity concentration, which determines the hole concentration, is kept low. In either case, fewer electrons will be available in the base to recombine with holes.

If the base were broad and its hole concentration were high, the greater proportion of the electrons that constitute the emitter current would, on diffusing across the base, recombine with holes and would fail to reach the collector junction. The collector current would not practically be augmented by the electrons supplied by the emitter, and only the base current would show some increase.

When no voltage is impressed on the emitter junction, practically no current is flowing across it. In the circumstances, the collector junction presents a high resistance to direct current because the majority carriers move away from the junction, and depletion layers are produced on either side of the boundary. Only a very small reverse current is then flowing across the collector junction due to the minority carriers in one region moving towards those from the other, that is, electrons from the *p*-region and holes from the *n*-region.

If, however, the impressed input voltage produces a substantial emitter current, the emitter will inject into the base a certain number of electrons which are minority carriers for that region. Since they have no time to recombine with holes as they diffuse through the base, they reach the collector junction. The greater the emitter current, the greater the number of electrons reaching the collector junction, and the lower its resistance. The collector current rises in proportion. In other words, a rise in the emitter current brings about an increase in the concentration of minority carriers injected from the emitter into the base. As a consequence, the increase in the number of these carriers leads to a rise in the collector current i_C.

As will be recalled, we use the term 'emitter' (which is short for 'emitter region') in order to stress the fact that it is responsible for the injection of electrons through the emitter junction into the base. The reference to 'injection' is important so that we could tell this process from electron emission as it occurs in a vacuum or a rarefied gas and produces free electrons.

To follow the above definition, we apply the name **'emitter'** to a transistor's region whose purpose is to inject carriers into the base. The name **'collector'** is assigned to the region whose purpose is to extract carriers from the base. The **base** is the region into which the emitter injects the carriers that are minority carriers for that region.

It should be noted that the emitter and the collector may exchange places (the inverted mode of operation). However, the collector junction of a transistor is, as a rule, larger in area than its emitter junction because it has to be able to dissipate substantially more power than the emitter junction. Therefore, if one chooses to use the emitter as a collector, the transistor would be operable, but at an appreciably lower power level, which is unattractive. If the two junctions are made the same in area (as is the case with *symmetrical transistors*), any of the outer regions may equally well be used as the emitter or as the collector.

Since in a transistor the emitter current is always the sum of the collector and base currents, the incremental change in emitter current must likewise be always equal to the sum of the incremental changes in collector and base currents:

$$\Delta i_E = \Delta i_C + \Delta i_B \tag{4-3}$$

An important property of transistors is that their currents are connected by an almost linear relation or, in simpler words, the three currents of a transistor vary in an approximate proportion to one another. Let us take as an example $i_E = 10$ mA, $i_C = 9.5$ mA, and $i_B = 0.5$ mA. If the emitter current goes up by, say, 20% to become equal to $10 + 2 = 12$ mA, the remaining two currents will rise likewise by 20% so that $i_B = 0.5 + 0.1 = 0.6$ mA and $i_C = 9.5 + 1.9 = = 11.4$ mA. This happens because the equality defined in Eq. (4-2) must always be satisfied, that is,

$$12 \text{ mA} = 11.4 \text{ mA} + 0.6 \text{ mA}$$

For the incremental changes in the currents, the equality defined in Eq. (4-3) holds, that is,

$$2 \text{ mA} = 1.9 \text{ mA} + 0.1 \text{ mA}$$

We have examined physical processes occurring in the *n-p-n* type of transistor. But everything said fully applies to the *p-n-p* type except that its electrons and holes exchange their roles and the polarity (sign) of the voltages and currents are reversed (Fig. 4-2*b*). In a *p-n-p* transistor, instead of electrons the emitter injects into the base holes which are minority carriers for that region. As the emitter current rises, a progressively greater number of such holes move through the base towards the collector junction. This brings down its resistance and raises the collector current.

The operation of a transistor can readily be visualized by reference to the potential diagram shown in Fig. 4-3 for an *n-p-n* transistor. This diagram can conveniently be used to build a mechanical model of a transistor. The emitter potential is taken as the datum or reference (zero) potential. There is a low potential barrier in the emitter junction. The greater the emitter-base voltage, v_{BE}, the lower this barrier. The potential difference across the collector junction is substantial, and it accelerates the electrons. In our mechanical model, balls replace electrons; owing to their inherent velocity, they climb the hill which models the emitter junction, pass through the region simulating the base of a transistor, and roll at an ever increasing velocity (that is, are accelerated) down the slope which simulates the collector junction.

In addition to the processes we have just discussed, some other events take place in a transistor that must be taken into account.

The performance of a transistor is materially affected by the base resistance r_{B0}*, that is the resistance that the base presents to the base current, i_B. This current flows to the base electrode in a direction which is at right angles to the direction from the emitter to the collector. Because the base is very thin, its resistance seen looking from emitter to collector, that is, to i_C, is very small and is usually neglected. In the direction of the base electrode, however, the base resistance, r_{B0}, runs into hundreds of ohms because in this direction the base acts similarly to a very fine conductor. The voltage across the emitter junction is always lower than v_{BE} because some of the applied voltage is dropped across the base resistance. If we take r_{B0} into account, a d. c. equivalent circuit for a transistor may be drawn up as shown in Fig. 4-4. In this diagram, r_{E0} is the emitter resistance which includes the resistance of the emitter junction and the resistance of the emitter region. Low-power transistors have r_{E0} running into tens of ohms. This is because the voltage across the emitter junction does not exceed a few tenths of a volt while the emitter current in such transistors is several milliamperes. Higher-power transistors have greater values of i_{E0} and proportionately smaller values of r_{E0}. Approximately, r_{E0} can be found (in ohms) by the following equation:

$$r_{E0} = 25/i_E \qquad (4\text{-}4)$$

where i_E is in milliamperes.

The collector resistance r_{C0} is practically that of the collector junction and may be as high as a few kilohms to tens of kilohms. It also includes the resistance of the collector region, but it is negligibly small.

The equivalent circuit in Fig. 4-4 is a very approximate one because in an actual transistor, the emitter, base and collector are in contact at a multiplicity of points over the entire area of the junctions. Still, this equivalent circuit may be used when examining many processes that occur in transistors.

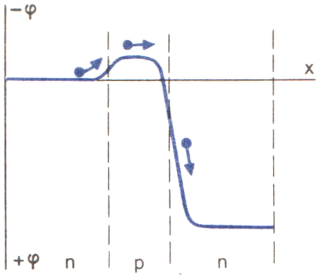


Fig. 4-3
Potential diagram of a transistor

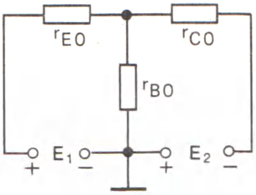


Fig. 4-4
D.C. equivalent circuit of a transistor

When the voltage applied to the collector junction is raised, what is known as the *avalanche multiplication of carriers* takes place, mainly due to impact ionization. This effect coupled with tunnelling may ultimately result in an electric breakdown. If the current is allowed to build up without bound, the electric breakdown may terminate in a thermal breakdown.

A change in the voltage existing across the collector and emitter junctions is accompanied by a change in the width of the junctions. In turn, this produces a change in the base width. This is known as *base width modulation.* It becomes a vital factor when there is an increase in the collector-base voltage – this leads to an increase in the width of the collector junction and a decrease in the base width. Should the base be too narrow, the collector junction may spread as far as the emitter junction and the two will make electrical contact with each other – it is then said that a *punch-through* (or a *reach-through*) has occurred. The base region then disappears altogether, and the transistor ceases operating as it should.

When the carrier injection from the emitter into the base is increased, the minority carrier

* Here and elsewhere, the zero (0) in a subscript indicates that the quantity involved is considered under conditions of direct-current flow.

concentration and charge in the base are increased, too, owing to what is known as *carrier storage*. Conversely, a reduction in the carrier injection will cause the stored charge to be removed from the base because both the minority carrier concentration and the minority carrier charge will then be brought down.

In some cases it is important to consider the flow of leakage currents over the surface of a transistor as it is accompanied by carrier recombination in the surface layer of all the regions of the device.

Let us establish the relations that exist between the currents flowing in a transistor. The emitter current is controlled by the voltage existing across the emitter junction, but the current reaching the collector is somewhat smaller in value – it may be called the *controlled collector current*, i_{Cc}. This happens because some of the carriers injected from the emitter into the base recombine. Therefore,

$$i_{Cc} = \alpha i_E \tag{4-5}$$

where α is the emitter-to-collector current gain of a transistor connected in a common-base circuit (also called the *alpha current factor* or the *common-base forward-current transfer ratio*). It may range in value from 0.950 to 0.998. That is, for a junction transistor it is always less than unity. It tends to approach unity with fewer injected carriers in the base recombining.

There is one more, very small current (not over a few microamperes) always flowing across the collector junction, symbolized as i_{CO} (Fig. 4-5). It is called the *reverse collector leakage current*. It is defined as the minimum current that will flow in the collector circuit of a transistor with zero current in the emitter circuit. Thus the total collector current is

$$i_C = \alpha i_E + i_{CO} \tag{4-6}$$

The reverse collector leakage current may be called an uncontrolled reverse current because it does not flow across the emitter junction.

In many cases, i_{CO} is a negligible fraction of i_E, so we may take it that i_C is approximately equal to αi_E. When measuring i_{CO}, the emitter circuit is opened. As follows from Eq. (4-6), when $i_E = 0$, $i_C = i_{CO}$.

Let us re-write Eq. (4-6) so that i_C is a function of i_B. On replacing i_E with the sum $i_C + i_B$, we get

$$i_C = \alpha(i_C + i_B) + i_{CO}$$

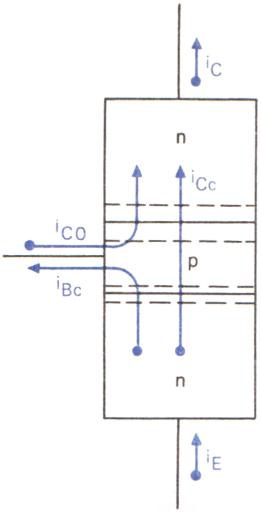


Fig. 4-5
Currents in a transistor

On solving the above equation for i_C, we obtain

$$i_C = i_B\,\alpha/(1-\alpha) + i_{CO}/(1-\alpha)$$

On denoting

$$\alpha/(1-\alpha) = \beta \quad \text{and} \quad i_{CO}/(1-\alpha) = i_{CEO}$$

we may finally write

$$i_C = \beta i_B + i_{CEO} \tag{4-7}$$

where β is the *beta current gain factor* or the *current transfer ratio* or *gain* of a transistor connected in a common-emitter circuit. It is always greater than unity and practical values up to 500 are often used.

It is interesting to compare how changes in α affect the value of β. For example, if $\alpha = 0.95$, then

$$\beta = \alpha/(1-\alpha) = 0.95/(1-0.95)$$
$$= 0.95/0.05 = 19$$

If $\alpha = 0.99$, which means an increase of 0.04 in its value, then

$$\beta = 0.99/(1-0.99) = 0.99/0.01 = 99$$

Thus, β increases more than five-fold. In other words, even minor changes in α lead to great changes in β. Similarly to α, β is a very important parameter of transistors. If we know β, we can always find α by the equation

$$\alpha = \beta/(1+\beta) \tag{4-8}$$

The current i_{CE0} is the reverse emitter current when the base is open, that is, when $i_B = 0$. It flows through all the three regions and the two junctions of a transistor. Thus, as follows from Eq. (4-7), if we set i_B to zero, we will get $i_C = i_{CE0}$. This current runs into tens or even hundreds of microamperes and greatly exceeds the reverse collector leakage current i_{C0}. More specifically,

$$i_{CE0} = i_{C0}/(1 - \alpha)$$

Therefore, if we know that $\alpha/(1 - \alpha) = \beta$, it is an easy matter to find that

$$i_{CE0} = (\beta + 1)\, i_{C0}$$

Also, since β is substantially greater than unity, we get

$$i_{CE0} \approx \beta i_{C0} \tag{4-9}$$

The high value of i_{CE0} is due to the fact that a small fraction of v_{CE} is applied to the emitter junction in the forward direction. In consequence, there is a rise in the emitter current which is, in this case, i_{CE0}.

When v_{CE} rises too high, i_{CE0} may go up abruptly and bring about an electric breakdown. If v_{CE} is not too low and the base circuit is open, a cumulative build-up of current may occur, leading to an excessive temperature rise and causing the transistor to fail (unless there is a current-limiting resistor in the collector lead). The events taking place may be summed up as follows. Some of the collector-to-emitter voltage, v_{CE}, causes a rise in i_E and in i_C which is equal to it, more carriers reach the collector junction, its resistance and the voltage across it go down, leading to a higher voltage across the emitter junction, and this in turn gives rise to a further increase in the current, and so on. To avoid this occurrence, the base circuit of a transistor may never be opened in operation, unless its collector supply voltage has been turned off. Also, the base supply voltage must be turned on first and the collector supply voltage afterwards – never do this the other way around.

When a need arises to measure i_{CE0}, place a current-limiting resistor in the collector lead and open the base circuit.

4-3 Amplification by a Transistor

Figure 4-6 shows the circuit diagram of an amplifier stage using an *n-p-n* transistor. This circuit is known as the *common-emitter* configuration (see Sec. 4-4) because the emitter is common to both the input and output circuits of the stage. The input voltage to be amplified is applied from a signal source, *SS*, across the emitter junction. The base also receives a positive bias voltage from a bias source E_1; this is the forward voltage for the emitter junction. As a result, a current is caused to flow in the base circuit and, in consequence, the input resistance of the transistor acquires a relatively low value. To prevent some of the input a. c. voltage from being dropped across the internal resistance of the signal source, the latter is shunted by a capacitor C_1 which has a relatively high capacitance. Its value is chosen so that its impedance is a very small fraction of the transistor's input resistance at the lowest operating frequency.

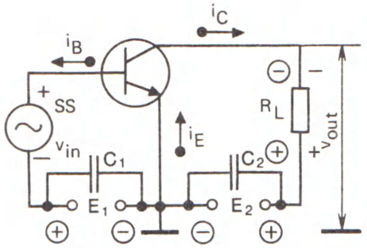


Fig. 4-6

Connection of a transistor in an amplifier stage (the common-emitter configuration)

The collector circuit (that is, the output circuit) is energized from another source, E_2. The amplified output voltage is taken off a load resistance, R_L. The E_2 source is shunted by a capacitor, C_2, so as to prevent some of the amplified voltage from being lost across the internal resistance of the E_2 source. At the lowest operating frequency, the impedance presented by this capacitor must be a very small fraction of R_L. In our further discussion we will omit the capacitors C_1 and C_2 of the E_1 and E_2 sources from the diagram so as to make it simpler to read. It may be presumed that they exist inside the E_1 and E_2 sources themselves. If these sources are rectifiers, they will always contain high-value capacitors in order to smoothen the pulsations.

A transistor amplifier stage operates as follows. Let the collector circuit be represented by the equivalent circuit of Fig. 4-7. The collector supply voltage E_2 is divided between the load resistance R_L and the internal resistance r_0 of the transistor which it presents to the direct collector current. This resistance is approximately equal to the resistance r_{C0} that the collector junction presents to direct current. Actually, to this resistance r_{C0} we ought to add the low resistances of the emitter junction and of the n- and p-regions, but they may be neglected.

When the signal source connected to the input circuit is turned on, its voltage causes the emitter current to vary. This brings about changes in the d.c. resistance of the collector junction, r_{C0}. In consequence, the collector supply voltage E_2 is re-distributed between R_L and r_{C0} in such a way that the a.c. voltage across the load resistor may be tens of times greater than the input a.c. voltage. Variations in the collector current are nearly equal to variations in the emitter current and are many times as great as the changes in the base current. For this reason the circuit we are discussing produces an appreciable current amplification and a very large amplification of power. The amplified power is part of the power supplied by the collector supply voltage E_2.

For better insight into the matter, we will trace the operation of the transistor amplifier stage by taking a numerical example. Let $E_1 = 0.2$ V, $E_2 = 12$ V, and $R_L = 4$ kΩ. Also let the d.c. transistor resistance r_0 in the no-signal condition be equal likewise to 4 kΩ, so that the total resistance of the collector circuit is 8 kΩ. Then the collector current which may approximately be taken equal to the emitter current will be

$$i_C = E_2/(R_L + r_0) = 12 \div 8 = 1.5 \text{ mA}$$

The collector supply voltage E_2 will be divided into two equal parts, and the voltages across R_L and r_0 will each be equal to 6 V.

If the signal source supplies an a.c. voltage with an amplitude of 0.1 V, the maximum voltage between the base and the emitter during the positive half-cycles will be 0.3 V. Suppose that this voltage causes the emitter current to rise to 2.5 mA. The collector current will practically rise to the same value. It will produce a voltage drop of 2.5 × 4 = 10 V across the load resistor, and the voltage drop across the transistor resistance r_0 will fall to 12 − 10 = 2 volts. In consequence, this resistance will fall to 2 ÷ 2.5 = 0.8 kΩ. During the negative half-cycles, when the signal source supplies a voltage of − 0.1 V, a reverse chain of events will take place. The minimum emitter-to-base voltage will become 0.2 − 0.1 = 0.1 V, the emitter and collector currents will fall to 0.5 mA each, the voltage drop across R_L will go down to 0.5 × × 4 = 2 V and that across r_0 will rise to 10 V, thereby implying that this resistance has risen to 10 ÷ 0.5 = 20 kΩ. Thus, feeding an a.c. voltage with an amplitude of 0.1 V to the transistor input causes r_0 to change from 0.8 to 20 kΩ, while the voltages across the load resistor and across the transistor change by 4 V either way (from 10 to 2 V). Therefore, the amplitude of the output voltage is 4 V which is 40 times the input voltage. (This is an approximate figure because actually a nonlinear relation exists between collector current and input voltage.)

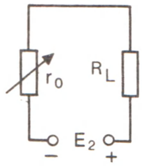


Fig. 4-7

Collector equivalent circuit of a transistor amplifier stage

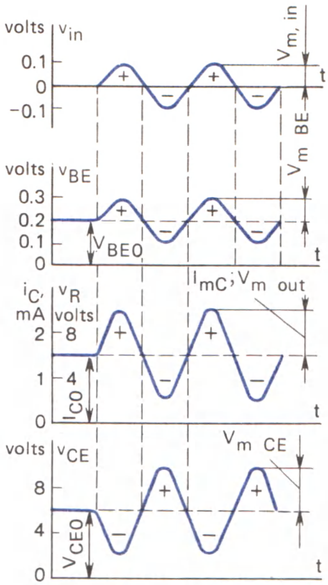


Fig. 4-8

Amplification by a transistor

Changes in the voltages and currents involved in our example are illustrated in the plots of Fig. 4-8. These plots correspond to the following equations:

- Input voltage:

$v_{\text{in}} = V_{m\,\text{in}} \sin \omega t$

- Base-emitter voltage:

$v_{BE} = \text{V}_{\text{BE}0} + V_{m\,\text{BE}} \sin \omega t$

where $V_{m\,BE} = V_{m\,\text{in}}$

- Collector current:

$i_{\text{C}} = I_{\text{C}0} + I_{m\,\text{C}} \sin \omega t$

- Load voltage:

$v_R = V_{R0} + V_{mR} \sin \omega t$

where $V_{mR} = V_{m\,\text{out}} = V_{m\,\text{CE}} = I_{m\,\text{C}} R_{\text{L}}$

$V_{R0} = I_{\text{C}0} R_{\text{L}}$

- Output voltage:

$v_{\text{out}} = v_{\text{CE}} = V_{\text{CE}0} - V_{m\,\text{CE}} \sin \omega t$

where $V_{\text{CE}0} = E_2 - V_{R0}$.

4-4 The Basic Circuit Configurations of Transistors

The three basic circuit configurations into which a transistor may be connected are the *common-emitter circuit*, the *common-base circuit*, and the *common-collector circuit*.

In each case one of the electrodes is common to the input and output of a stage. To avoid confusion, it should be remembered that we use the terms 'input' and 'output' as referring to the points between which the input and output a. c. voltages exist, and not the d. c. supply voltages used by the device.

The three circuit configurations have the following abbreviations: CE, CB, and CC. Sometimes, the term "common" is replaced with the term "grounded" (for example, a grounded-base circuit), although grounding will not always be provided.

The three types of circuit amplify signals by the same principle, of course, but they differ in some properties and so they should be discussed in detail separately.

The common-emitter (CE) connection. This configuration shown in Fig. 4-6, is the most commonly used one because it gives the largest power amplification of all.

The current gain of such a stage, k_I, is the ratio between the peak (or rms) values of output and input alternating currents, that is, the a. c. components of collector and base currents:

$$k_I = I_{m\,\text{out}}/I_{m\,\text{in}} = I_{m\text{C}}/I_{m\text{B}} \qquad (4\text{-}10)$$

Because the collector current is tens of times as great as the base current, k_I has a value of several tens.

The amplifying properties of a transistor with the common-emitter connection are stated in terms of one of transistors' principal parameters – the *beta current gain factor*, β, also known as the *static common-emitter forward-current transfer ratio*. Because it characterizes only the transistor itself, it must be measured with the load disconnected from the amplifier stage ($R_{\text{L}} = 0$), that is, at a static (direct) collector-emitter voltage:

$$\beta = \Delta i_{\text{C}}/\Delta i_{\text{B}} \text{ with } v_{\text{CE}} \text{ held constant} \qquad (4\text{-}11)$$

As already noted, β may practically be as high as several hundreds. The stage current gain k_I, however, is always smaller than β because bringing the load resistance R_{L} into circuit reduces the collector current, i_{C}.

The stage voltage gain, k_V, is defined as the ratio between the peak or rms values of output and input a. c. (or signal) voltages. The input signal is that between base and emitter, v_{BE}; and the output signal is the alternating voltage appearing across the load resistor, v_R, or, which is the same, that between collector and emitter, v_{CE}:

$$k_V = V_{m\,\text{out}}/V_{m\,\text{in}} = V_{mR}/V_{m\text{BE}} = V_{m\text{CE}}/V_{m\text{BE}} \qquad (4\text{-}12)$$

The emitter-to-base voltage does not exceed a few tenths of a volt, while the output voltage across a sufficiently high load resistance and a sufficiently high E_2 supply voltage may run into units of volts or even higher. Therefore, k_V ranges from tens to several hundreds.

It follows, therefore, that the stage power gain, k_P, may range from several hundreds to tens of thousands. This gain factor is calculated by dividing the output power by the input power. Each is equal to half the product between the amplitudes of the respective currents and voltages:

$$P_{\text{out}} = I_{m\,\text{out}} V_{m\,\text{out}}/2 = I_{m\text{C}} V_{m\text{CE}}/2 \qquad (4\text{-}13)$$

$$P_{\text{in}} = I_{m\,\text{in}} V_{m\,\text{in}}/2 = I_{m\text{B}} V_{m\text{BE}}/2 \qquad (4\text{-}14)$$

Therefore,

$$k_P = P_{\text{out}}/P_{\text{in}} = I_{m\,\text{out}} V_{m\,\text{out}}/I_{m\,\text{in}} V_{m\,\text{in}} = k_I k_V \qquad (4\text{-}15)$$

An important quantity characterizing a transistor is its *input resistance* R_{in} which is found by Ohm's law. For a CE connection it is

$$R_{\text{in}} = V_{m\,\text{in}}/I_{m\,\text{in}} = V_{m\text{BE}}/I_{m\text{B}} \qquad (4\text{-}16)$$

and ranges from hundreds of ohms to several kilohms.

This stems from the fact that with $V_{m\text{BE}}$ equal to a few tenths of a volt the base current $I_{m\text{B}}$ of low- and medium-power transistors may be as small as a few tenths of a milliampere. For example, if $V_{m\text{BE}} = 200$ mV and $I_{m\text{B}} = 0.4$ mA, then $R_{\text{in}} = 200 \div 0.4 = 500\ \Omega$. As is seen, the input resistance is relatively small. This is a major disadvantage of bipolar transistors. As will be shown later, the output resistance of a transistor connected in a common-emitter circuit is from units to tens of kilohms.

A CE transistor amplifier stage inverts the phase of the signal voltage – the phase difference between the input and output voltages is 180°. To prove this, let us see how the circuit shown in Fig. 4-6 operates. In this and the subsequent figures the polarities of the d.c. potentials are labelled with an encircled plus sign so that they could not be confused with alternating potentials. The direct collector current produces across the load resistor a voltage drop such that the " – " terminal is at the top end of the resistor (as it appears in the figure). Let the transistor's input (base) accept a positive half-cycle of voltage as shown in Fig. 4-6. This voltage is added to E_1, and the voltage across the emitter junction, V_{BE}, goes up. This brings about an increase in the emitter current and, as a consequence, in the collector current. As a result, the voltage drop across the load resistor is increased because in addition to a d.c. voltage across R_{L} there appears an a.c. voltage of the same polarity. In this way, a negative half-cycle of voltage appears at the output.

An advantage of the common-emitter connection is that it needs only one supply source because the collector and the base are energized with supply voltages of the same polarity.

A drawback of the CE connection in comparison with the common-base (CB) circuit is an impairment in the frequency and temperature properties. As the signal goes up in frequency, the gain of the CE circuit falls off in a greater proportion that it does in the CB circuit. The operating currents and voltages of the CE circuit are heavily dependent on temperature. The effects of frequency and temperature will be examined in more detail in Chap. 6.

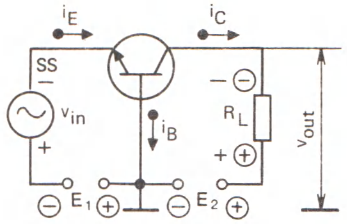


Fig. 4-9

Connection of a transistor in a common-base (CB) circuit

The common-base (CB) connection. Although this circuit configuration (Fig. 4-9) yields a substantially lower power gain and has a still lower input resistance than the CE connection, it is used sometimes all the same because it compares favourably with the CE connection with regard to frequency and temperature (see Chap. 6).

The current gain of a CB stage is always slightly less than unity:

$$k_I = I_{m\text{C}}/I_{m\text{E}} \approx 1 \qquad (4\text{-}17)$$

because the collector current is always only slightly smaller than the emitter current.

We have already defined the common-base forward-current transfer ratio of a transistor, also known as the *static alpha current gain factor*, α. We have also stressed that the alpha current gain factor (or the static current gain) of a transistor connected as a CB amplifier is measured with its load disconnected, $R_{\text{L}} = 0$, that is, with the collector-to-base voltage held constant:

$$\alpha = \Delta i_{\text{C}}/\Delta i_{\text{E}} \text{ with } v_{\text{CB}} \text{ held constant} \qquad (4\text{-}18)$$

We have also noted that the alpha current gain factor is always less than unity and that the closer it is to unity, the better the performance of the transistor. The stage current gain k_I of the common-base connection is always somewhat smaller than the alpha current gain factor because bringing R_{L} into circuit reduces the collector current.

The voltage gain of a CB stage is defined as

$$k_V = V_{m\text{CB}}/V_{m\text{EB}} \qquad (4\text{-}19)$$

It is the same in value as in the CE connection, that is, it ranges from tens to hundreds. This cannot be otherwise because if the CE circuit and the CB circuit use the same type of transistors, the same input (signal) and supply voltages, and the same load resistances, the collector current will likewise be practically the same, and so the output voltages will be the same too. Because the power gain k_P is the product of current and voltage gains, $k_I k_V$, and k_I is approximately equal to unity, it follows that k_P is about the same as k_V, that is, its value ranges from tens to hundreds.

The input resistance of a CB amplifier stage is

$$R_{\text{in}} = V_{m\text{EB}}/I_{m\text{E}} \tag{4-20}$$

It is by a factor of several tens smaller than in the CE connection. This naturally stems from the fact that $V_{m\text{EB}}$ is equal to $V_{m\text{BE}}$ while $I_{m\text{E}}$ is tens of times the value of $I_{m\text{B}}$. For the CB connection, R_{in} is only several tens of ohms while for CB circuits using high-power transistors it is even a few ohms. Such a small input resistance is a major limitation of the CB configuration. Its output resistance, as will be shown later, may be several hundred kilohms.

The CB circuit does not reverse the phase of the input voltage, so there is no phase difference between the input and output voltages. This can be proved by reasoning along the same lines as in the case of the CE connection. Figure 4-9 shows that during the negative half-cycles of input (signal) voltage the emitter and collector currents increase in value, giving rise to a greater voltage drop across the load resistance, that is, to a higher negative half-cycle of output voltage.

It is worth while mentioning that a CB amplifier corrupts the signal substantially less than a CE amplifier.

The common-collector (CC) connection. In this configuration (Fig. 4-10) the collector is a true common point for the input and output circuits because the E_1 and E_2 sources are always shunted by high-value capacitors and it is legitimate to visualize them as virtual short-circuits for alternating current. A distinction of this configuration is that all of the output voltage is fed back to the input – it is said that a very large amount of negative feedback is used. It is an easy matter to see that the input voltage is the sum of the alternating base-to-emitter voltage, v_{BE}, and the output voltage:

$$v_{\text{in}} = v_{\text{BE}} + v_{\text{out}} \tag{4-21}$$

The current gain of a CC circuit is about the same as that of the CE connection – it has a value of several tens. To demonstrate,

$$k_I = I_{m\text{E}}/I_{m\text{B}} = (I_{m\text{C}} + I_{m\text{B}})/I_{m\text{B}} = I_{m\text{C}}/I_{m\text{B}} + 1 \tag{4-22}$$

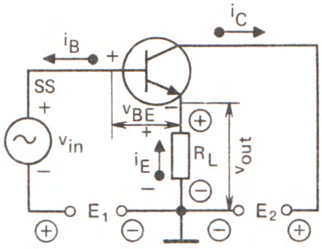


Fig. 4-10

Connection of a transistor in common-collector (CC) circuit

and the ratio $I_{m\text{C}}/I_{m\text{B}}$ is the common-emitter forward-current transfer ratio or current gain.

In contrast, the voltage gain is very close to, but always less than, unity:

$$k_V = V_{m\,\text{out}}/V_{m\,\text{in}} = V_{m\,\text{out}}/(V_{m\text{BE}} + V_{m\,\text{out}}) < 1 \tag{4-23}$$

The value of $V_{m\text{BE}}$ is a few tenths of a volt, and that of $V_{m\,\text{out}}$ is several volts so that $V_{m\text{BE}} \ll$ $\ll V_{m\,\text{out}}$. In consequence, k_V is approximately equal to unity.

It is to be noted that the alternating (signal) voltage applied to the input of the transistor is amplified tens of times, as is the case with the CE connection, but the stage as a whole does not produce amplification. Obviously, the power gain is about equal to k_I, that is, to several tens.

If we look at the polarities of the a. c. voltages existing in the CC circuit, we will see that no phase shift is introduced between v_{out} and v_{in}. Let, for example, the CC stage accept at some instant of time a positive half-cycle of input voltage v_{in}, as shown in Fig. 4-10. This will bring about a rise in v_{BE} and in the emitter current, with the latter leading to a greater voltage drop across the load resistor. In consequence, the output of the stage will produce a positive half-cycle of voltage which is in phase with the input voltage and nearly equal in magnitude to it. In other words, the output voltage follows the input voltage. Quite aptly this circuit configuration is alternatively called the *emitter follower*. The term 'emitter' is included in the name of the stage because the load resistor is placed in the emitter lead and the output voltage is picked off the emitter (with

Table 4.1

LEADING PARTICULARS OF THE BASIC TRANSISTOR CONNECTIONS

Parameter	CE	CB	CC
k_I	Tens to hundreds	Slightly less than unity	Tens to hundreds
k_V	Tens to hundreds	Tens to hundreds	Slightly less than unity
k_P	Hundreds to tens of thousands	Tens to hundreds	Tens to hundreds
R_{in}	Hundreds of ohms to several kilohms	Units to tens of ohms	Tens to hundreds of kilohms
R_{out}	Several to tens of kilohms	Hundreds of kilohms to several megohms	Hundreds of ohms to several kilohms
Phase difference between V_{out} and V_{in}	180°	0	0

respect to chassis ground).

The input resistance of a CC amplifier stage is tens of kilohms, and this is an important advantage of this circuit configuration. Thus,

$$R_{\text{in}} = V_{m\,\text{in}}/I_{m\,\text{in}} = (V_{m\text{BE}} + V_{m\,\text{out}})/I_{m\text{B}} \qquad (4\text{-}24)$$

The ratio $V_{m\text{BE}}/I_{m\text{B}}$ is the input resistance of the transistor itself connected in a common-emitter circuit, which, as has been noted, is several kilohms. Since, however, $V_{m\,\text{out}}$ is tens of times the value of $V_{m\text{BE}}$, it follows that R_{in} is tens of times the value of the input resistance in the CE circuit. On the contrary, the output resistance of a CC stage is relatively small, usually several kilohms or even hundreds of ohms.

Because the CC connection is not a very common occurrence, we will limit our further discussion to the CE and CB connections.

For ease of comparison, the basic properties of the three circuit configurations for bipolar transistors are summarized in Table 4-1.

4-5 Bias Supply and Temperature Compensation for Transistors

It is usual to employ a single source, that of the output circuit,* to energize a transistor stage. For the transistor to operate normally, it is essential that a d. c. voltage of a few tenths of a volt, known as the *base bias voltage*, be maintained between the emitter and base.

On passing across the emitter junction, the emitter current produces a voltage drop across it, but this drop is not enough to make the transistor operate normally. Therefore, unless an additional bias voltage is applied, the currents will be too small. This additional bias voltage is tapped from the collector supply source with the aid of a resistor or a divider. Figure 4-11 illustrates how the bias voltage is derived in several typical cases.

In a CE stage (Fig. 4-11*a*), the direct base current I_{B0} flows through a resistor R across which nearly all of the E_2 supply voltage is dropped. Only a very small proportion of E_2 is dropped across the emitter junction and serves as the base bias voltage:

$$V_{\text{BE0}} = E_2 - I_{\text{B0}}R \qquad (4\text{-}25)$$

It is an easy matter to derive the value of R from the above equation:

$$R = (E_2 - V_{\text{BE0}})/I_{\text{B0}} \qquad (4\text{-}26)$$

As a rule, $V_{\text{BE0}} \ll E_2$, and R is approximately equal to E_2/I_{B0}.

Figure 4-11*b* shows how the bias voltage is derived with the aid of a divider R_1R_2 in a CE stage. Here, the greater proportion of E_2 is dropped across R_1, and the smaller part, serving as the bias voltage V_{BE0} is dropped across R_2

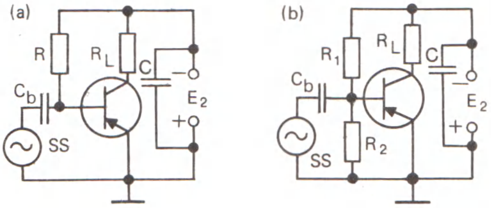


Fig. 4-11

Base biasing circuits of a transistor

[*] As agreed earlier, this source will be designated as E_2.

which is placed in shunt with the transistor input. The values of R_1 and R_2 can readily be calculated by the equations

$$R_1 = (E_2 - V_{BE0})/(I_{dv} + I_{B0}) \approx E_2/(I_{dv} + I_{B0})$$
$$R_2 = V_{BE0}/I_{dv} \qquad (4\text{-}27)$$

where I_{dv} is the current flowing through the divider.

It is a fairly common practice to derive voltage bias by means of a divider, but this method is wasteful of power because the source has to sustain an additional current, I_{dv}, which is uselessly dissipated as heat in R_1 and R_2. Also, there is a marked decrease in the input resistance of the stage because in the circuit in question R_2 is placed in shunt with the transistor input.

For bias voltage to be more stable, it is desirable that the divider current I_{dv} be as high as practicable. Then the division of voltage among the divider resistors will depend only slightly on the base current flowing through one of them. To save the power supplied by the source, however, it is usual to set I_{dv} at a value which is only 3 to 5 times the value of I_{B0}. The d.c. blocking capacitor C_b serves to pass the a.c. voltage to be amplified to the transistor input. The capacitance of C_b at the lowest operating frequency is chosen to be sufficiently small so as to minimize the voltage drop across the blocking capacitor. More often, C_b is chosen to have a capacitance of several microfarads or tens of microfarads. For this reason, C_b in low-frequency circuits is usually a small-sized electrolytic capacitor. In the circuits of Fig. 4-11*a* and *b*, the d.c. blocking capacitor, as its name implies, blocks the passage of direct current to the transistor input, if this is present in the signal voltage. There is another reason for provision of C_b. Without it and with a signal source of low internal resistance, the base and emitter would be short-circuited for d.c., and V_{BE0} would be close to zero. The value of C_b is calculated proceeding from the requirement that the reactance $1/\omega C_b$ at the lowest operating frequency must be a small fraction of the input resistance of the transistor:

$$1/\omega_l C_b \ll R_{in} \qquad (4\text{-}28)$$

Then the voltage drop across C_b will be kept to a low level. In practice it will suffice to satisfy the following condition:

$$1/\omega_l C_b \leqslant 0.1\, R_{in} \qquad (4\text{-}29)$$

Hence,

$$C_b \geqslant 10/\omega_l R_{in} \qquad (4\text{-}30)$$

If C_b is in microfarads, the design equation will have the form

$$C_b \geqslant 10 \times 10^6/2\pi f_l R_{in} \qquad (4\text{-}31)$$

Here, ω_l and f_l refer to the lowest operating frequency in angular units and in hertz, respectively.

For a circuit using a divider, the R_{in} of the transistor in the equation must be replaced with a resistance R'_{in} which is equivalent to R_{in} and R_2 connected in parallel, that is,

$$R'_{in} = R_{in}R_2/(R_{in} + R_2) \qquad (4\text{-}32)$$

The capacitance of the capacitor C placed in shunt with the E_2 source should be found by an equation similar to (4.31):

$$C \geqslant 10 \times 10^6/2\pi f_l R_L \qquad (4\text{-}33)$$

where f_l is the lowest operating frequency and R_L is the load resistance. If the condition defined in Eq. (4-33) is satisfied, practically all of the output voltage will be developed across R_L, and its fraction lost across E_2 will be small.

An important limitation of transistors is a marked change in their characteristics with temperature. An increase in temperature brings about a rise in currents so that the operating conditions of the transistors become other than normal. This undesirable occurrence can be controlled by any one of several methods whose objective is to compensate for the temperature effect and to stabilize the operating conditions. The circuit using a transistor (or transistors) is extended to include compensating elements or circuits so that the operating conditions could be maintained more or less constant in the face of variations in ambient temperature or secondary to the replacement of the transistor. It should be remembered, however, that these compensation circuits can only stabilize the operating point but cannot cancel the effect of temperature on the properties of a transistor and on the processes that take place in it. Therefore, changes in temperature do bring about changes in transistor parameters. Thus, we can remove the detrimental consequences of the temperature effect only in part.

Figure 4-12 shows the most commonly used simple compensation schemes for the CE configuration which is sensible to variations in temperatue most of all (see Chap. 6). In what is known as the *collector temperature-stabilized* (or *temperature-compensated*) *scheme*, the resistor R intended to provide the necessary base bias is connected to the collector rather than to the E_2 source as in the circuit of Fig. 4-11*a*. If heat build-up or a transistor replacement should cause a rise in i_C, the voltage drop across R_L will also increase and the voltage V_{CE} will fall in proportion. This will however bring about a decrease in V_{BE}, and this will lead to a decrease in i_C. In this way, the collector current is simultaneously changed in opposite senses so that its net value remains nearly constant.

The form of temperature compensation we have just discussed is simple and economical but its efficacy is good only if the voltage drop across the load resistor accounts for not less than half of the supply voltage E_2. Also, this scheme entails a degradation in amplification because some of the amplified voltage is fed back via the resistor R to the transistor input in anti-phase with the signal voltage to be amplified. This is in effect negative voltage feedback.

The scheme shown in Fig. 4-12*b* is more elaborate and more wasteful of power. It may be called the *emitter temperature-compensated* (or *temperature-stabilized*) *scheme*. With it, the E_2 source must supply a somewhat greater voltage but in terms of compensation it is superior to the previous arrangement. Here the resistors R_1 and R_2 make up a voltage divider which provides the required base bias, and the resistor R_E placed in the emmiter lead operates as the compensating element. The voltage drop $V_E = I_{E0} R_E$ across this resistor opposes the output voltage $V_2 = I_{dv} R_2$. Therefore, the base bias voltage is $V_{BE0} = V_2 - V_E$. The resistor R_E sets up negative feedback in terms of direct current. Should a rise in temperature cause the currents in the transistor to rise, the increase in I_{E0} will raise the emitter voltage V_E, and lower the base bias voltage V_{BE0} in proportion, thus leading to a fall in the currents. As a result of this simultaneous but opposite change, the net currents will change very little and the operating conditions will be made more stable.

The resistor R_E is placed in shunt with a sufficiently high-valued capacitor C_E so as to prevent it from setting up negative feedback in terms of alternating current. The reactance of this capacitor at the lowest frequency must be a very small fraction of R_E. As a rule, C_E is an electrolytic capacitor with a value of tens of microfarads (in a. f. amplifier stages). The emitter compensation scheme is effective irrespective of R_L, and its performance improves with increasing divider current I_{dv} and increasing emitter resistor R_E. Since, however, V_E is part of E_2, an excessive increase in R_E would cause one to use a higher E_2, which is a disadvantage. Neglecting V_{BE0} in comparison with the other voltages, the resistance values for the emitter compensation scheme can be calculated, using the following approximate equations:

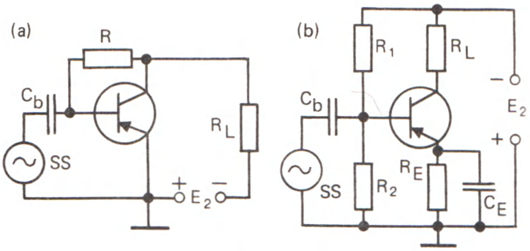


Fig. 4-12

Temperature compensation schemes for a transistor amplifier stage

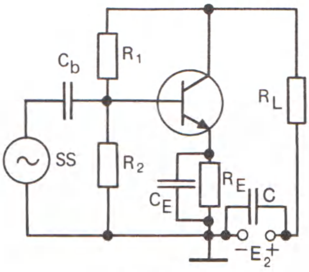


Fig. 4-13

Collector-emitter temperature-compensation scheme

$$R_1 \approx (E_2 - V_E)/(I_{B0} + I_{dv})$$

$$R_2 \approx V_E/I_{dv} \qquad (4\text{-}34)$$

$$R_E = V_E/I_{E0}$$

A further requirement is to choose the value of V_E in view of the likely increase in E_2. The divider current is usually anywhere between three to five times I_{B0}.

As an alternative, the two temperature compensation schemes may be used together – this will improve compensation still more (Fig. 4-13).

In many cases, temperature compensation is not mandatory because a highly stable gain is not essential.

A single supply source and temperature compensation may be used also when transistors are connected in CB or CC circuits.

4-6 Transistors in Amplifiers and Oscillators

We discuss simple circuits which illustrate the use of transistors in amplifiers and oscillators.

A transistor amplifier stage may have its output connected to any one of several types of load. If the load is a resistor, as has been shown in the previously discussed cases, we have what is called a *resistance-coupled stage*. At low frequencies, it is customary to use *transformer-coupled* stages (Fig. 4-14) in which the output circuit is coupled to the input of the next stage or some other load (say, a speaker) via the secondary of a transformer. Transformer-coupled stages may well be used at r.f., the load often being a resonant circuit tuned to the operating frequency (Fig. 4-15). This tuned circuit passes the amplified signal on to, say, the next stage. Not infrequently, the signal source in r.f. amplifier stages is likewise an input resonant circuit.

In the above and subsequent cases, the base bias voltage is fed from the E_2 source via a resistor R, and C_b is the d.c. blocking capacitor. Its function has been described already.

A signal source may sometimes be connected via an input transformer rather than via a capacitor (Fig. 4-16). In this arrangement, the function of C_b is to feed the a.c. voltage to the transistor input without losses across R.

An important application is the use of transistors in oscillators. A simple transistor oscillator employing inductive feedback is shown in Fig. 4-17. It operates as follows. When the supply voltage is turned on, a current begins to flow in the collector circuit and gives rise to free oscillations in the tuned (tank) circuit LC. Without a transistor, these oscillations would have died out because of power losses in the LC tank. Owing to the feedback coil L_1 which is inductively coupled to the tuned-circuit (or tank) coil L, the tank oscillations are fed back (via C_b) to the transistor input. If, as a result of amplification, the oscillations produced in the tank circuit are in phase with original free oscillations and are sufficiently strong to make up for the power lost in the tank, no decay will take place. The tank will sustain *undamped oscillations*. For the amplified oscillations to have the right phase and to sustain rather than dampen oscillations in the LC tank, a proper point of connection must be found for the feedback coil L_1.

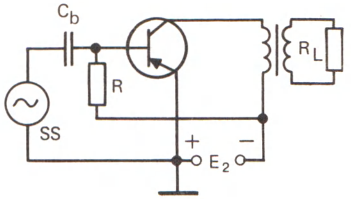


Fig. 4-14

Circuit diagram of a transformer-coupled amplifier stage

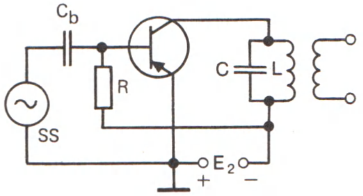


Fig. 4-15

Circuit diagram of an amplifier stage containing a resonant circuit

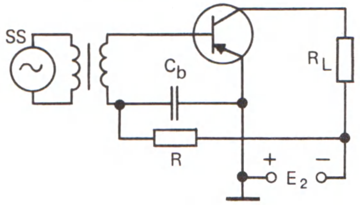


Fig. 4-16

Connection of a signal source to the input of a transistor via a transformer

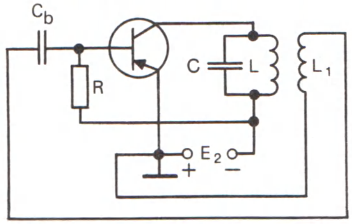


Fig. 4-17

Simple transistor oscillator using inductive feedback

In our examples, we have used the common-emitter connection for transistors. Of course, other circuit configurations may also be used.

Chapter Five
Characteristics and Parameters of Bipolar Transistors

5-1 Characteristics of Transistors

The relations between the currents obtained in and the voltages applied to a transistor are graphically presented as characteristic curves. When they are plotted for direct currents only and with no load connected to the output, they are called *static characteristics*. When they are plotted for a. c. and with some load connected to the output, we have *dynamic characteristics*. Characteristics are an important tool in investigating the properties of transistors and in practical calculations of transistor circuits.

Four quantities are always interrelated in a transistor. These quantities are the input and output currents and voltages, i_1, i_2, v_1, and v_2. A single family of characteristic curves would be unable to depict all the likely relations. Instead, use is made of two families of curves. It is most convenient to deal with a family of *input characteristics* $i_1 = f(v_1)$ along with a family of *output characteristics* $i_2 = f(v_2)$.

Each of the three basic circuit configurations we have examined previously has families of characteristics of its own. Therefore, when using characteristics, it is important to note to which circuit configuration they apply. We will examine the characteristics applicable to the CE and CB configurations as they are most commonly used. These characteristics are usually given in data sheets and other reference sources.

Sometimes characteristic curves are plotted considering the fact that the voltages applied to and the currents obtained in *n-p-n* and *p-n-p* transistors differ in polarity (sign), that is, with negative voltages and currents laid off to the left and downwards from the origin. In many cases, however, it is convenient to lay them off to the right and upwards of the origin. Exactly such characteristics will be used in our further discussions. The polarity of voltages applied to a transistor and the direction of current flow in its circuits are always established according to the type of transistor used and irrespectively of the manner in which its characteristics are constructed.

The input and output characteristics of a transistor are similar to the characteristics of a semiconductor diode. The point is that the input characteristics hold for the emitter junction which is always forward-biased. Therefore, they are analogous to the characteristic of a forward-biased diode. The output characteristics of a transistor are similar to the characteristic of a reverse-biased diode because they reflect the properties of the collector junction when it is reverse-biased.

To begin with, we will examine the characteristics of a transistor connected in a CE circuit.

Figure 5-1*a* shows the input characteristics $i_B = f(v_{BE})$ with the output voltage held constant (v_{CE} = const). At $v_{CE} = 0$, the characteristic starts at the origin because, if all voltages are zero, there can be no currents flowing.

As is seen from Fig. 5-1*b*, when $v_{CE} = 0$, that is, when the collector and emitter are short-circuited, a forward voltage equal to $v_{BE} = E_1$ is applied to both junctions. In the circumstances, the base current is the sum of the forward currents crossing the emitter and collector junctions, but it is small because the forward voltage applied to the emitter junction is just a few tenths of a volt (several hundred millivolts), and the series base resistance r_{B0} is hundreds of ohms.

In low-power transistors, the base current is not over several tens or hundreds of micro-

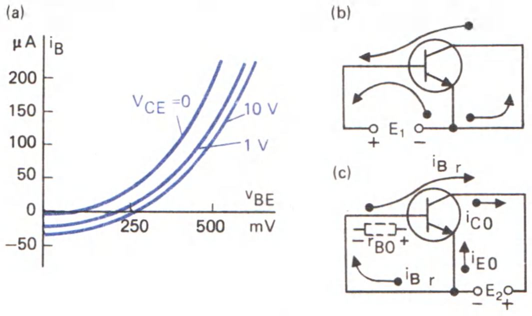


Fig. 5-1

Input characteristics of a transistor in a CE circuit amperes. The characteristic we are discussing is similar to the usual characteristic of a forward-biased semiconductor diode. At $v_{CE} > 0$, the curve is shifted to the right, the base current is reduced and becomes negative at low values of v_{BE}. This is illustrated in Fig. 5-1*c* which shows a CE circuit with $v_{BE} = 0$, that is, with E_1 disconnected. In this case, the E_2 source produces both the emitter leakage current i_{E0} and the reverse base current i_{Br}. On combining, the two currents produce the reverse collector leakage current, $i_{C0} = i_{E0} + i_{Br}$. It is to be noted that i_{Br} produces across r_{B0} a small voltage drop which is a forward-bias voltage for the emitter, and so it somewhat raises the reverse emitter leakage current i_{E0}. If we now turn on the E_1 source and gradually raise its voltage, it will oppose the action of E_2 in the base circuit. This will reduce i_{Br} until it falls to zero at some value of v_{BE} when E_1 and E_2 balance each other. A further rise in v_{BE} will build up the positive base current which is, as a rule, part of the emitter current.

Another cause for a reduction in base current with a rise in v_{CE} is base width modulation. The higher the value of v_{CE}, the greater the voltage across the collector junction, v_{CB}. The width of the collector junction is increased and that of the base is decreased, and fewer carriers moving from emitter to collector recombine in the base. In consequence, i_C rises and i_B falls. However, a change in v_{CE} (from, say, 1 to 10 V, as is shown in Fig. 5-1*a*) affects the base current but little. The input characteristics plotted for several values of v_{CE} are closely spaced. Reference sources and data sheets usually give only one input characteristic for the recommended value of v_{CE}. As often as not, they give also a characteristic for $v_{CE} = 0$.

A family of output characteristics $i_C = f(v_{CE})$ is shown in Fig. 5-2*a*. As a rule, these characteristics are plotted for several values of direct base current. The explanation is that the input resistance of a transistor is relatively small and the source of the input a.c. voltage (the signal source), often having a high internal resistance, operates as an equivalent constant-current generator (the Norton equivalent). Therefore, it is usual to specify the input current of a transistor, so that calculations can conveniently be made by reference to a family of output characteristics relating the output current and voltage to the input current.

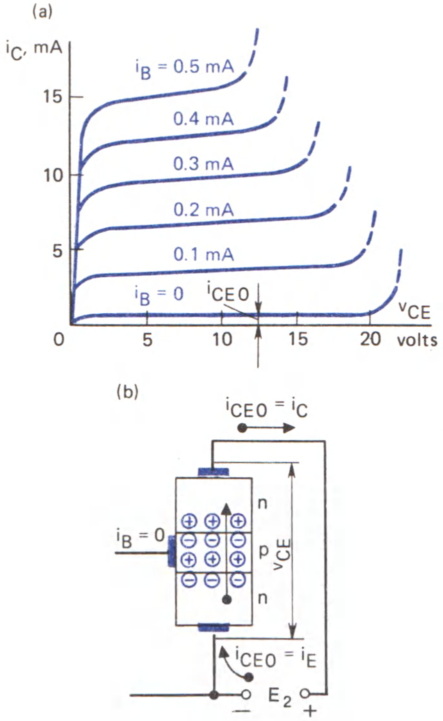


Fig. 5-2

Output characteristics of a transistor in a CE circuit

The first curve plotted for $i_B = 0$ starts at the origin and looks very like the usual characteristic of a reverse-biased semiconductor diode. The equality $i_B = 0$ implies that the base circuit is open. In the circumstances, what may be called the *transfer collector-to-emitter current*, i_{CEO} (Fig. 5-2*b*), flows through the entire transistor.

When $i_B > 0$, the output characteristic is shifted upwards and runs higher than it does at $i_B = 0$. This upward shift increases with increasing i_B. The increase in the base current signifies that a rise in V_{BE} brings about an increase in the emitter current which includes as one of its components the base current i_B. As a result, there is a proportionate increase in the collector current as well. Owing to a linear relationship between the currents, the quietly sloping portions of adjacent characteristic curves are spaced an about equal distance apart. In some transistors, however, this linearity is somewhat disturbed.

Output characteristics show that as v_{CE} rises from zero to some low values (a few tenths of a volt), the collector current builds up at a very high rate while with any further increase in v_{CE} the curves have a small up-slope. This points to the fact that v_{CE} only slightly affects the collector current. For a large increase in i_C, one must increase the emitter current. Still, the following happens when v_{CE} is raised. Because the base width is reduced, the base current falls and, since the characteristics are plotted at $i_B =$ const, one has to raise v_{BE} so as to hold the base current at its previous value. Because of this there is a rise in i_E and, as a consequence, in the collector current. A rise in v_{CE} also entails a rise in its component that is applied as a forward-bias voltage to the emitter junction. As a result, i_E and i_C also increase.

The characteristics in Fig. 5-2*a* show that high values of collector current lead to an earlier electrical breakdown, that is, a breakdown occurs at lower values of v_{CE}. The breakdown region is not the normal operating region for a transistor.

Sometimes, use is made of the output characteristics $i_C = f(v_{CE})$ plotted for several values of direct emitter-to-base voltage, v_{BE}. It is convenient to use them when the input voltage is known or specified in advance and the signal source has a low internal resistance (a small fraction of the input resistance of the transistor) so, that the source may be treated as an equivalent constant-voltage generator (the Thévenin equivalent). Such a family of curves is shown in Fig. 5-3. Their distinction is that the various curves are spaced different distances apart. At low values of v_{BE} the curves run closer together. This is an outcome of the nonlinear relation between i_C and v_{BE}. As will be recalled, i_C is approximately proportional to i_B, but i_B is a nonlinear function of v_{BE}, which is clearly seen from the input characteristic in Fig. 5-1*a*. When $v_{BE} = 0$ the collector circuit carries a small reverse collector leakage current i_{C0} mentioned earlier. A substantial rise in v_{CE} may cause an electrical breakdown.

Although it will usually suffice to have the input and output characteristics of a transistor in order to design and calculate the circuit using it, sometimes resort is also made to *control* (or *transfer*) *characteristics* which relate i_C to i_B with v_{CE} held constant (Fig. 5-4*a*) or $i_C = f(v_{BE})$ with v_{CE} held constant (Fig. 5-4*b*). These characteristics clearly show that the relation between i_C and i_B is linear very nearly while the relation between i_C and the input voltage is nonlinear.

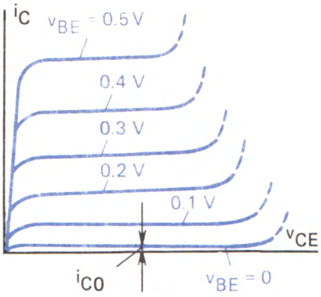


Fig. 5-3

Output characteristics of a transistor with its emitter-to-base voltage held constant

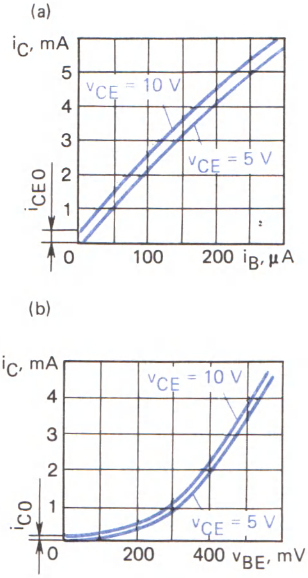


Fig. 5-4

Transfer characteristics of a transistor connected in a CE circuit

Changes in v_{CE} affect i_C only slightly, and the control (transfer) characteristics for several values of v_{CE} run very closely to one another. Reference sources and data sheets usually give only one curve for some average value of v_{CE}. When $i_B = 0$, a small collector-to-emitter (or transfer) current is flowing. When $v_{BE} = 0$, there is a small reverse collector leakage current, i_{C0}, flowing.

As we have seen, i_C and i_B are connected by a relation of the form

$$i_C = \beta i_B + i_{CE0} \tag{5-1}$$

If we assume that β is constant, Eq. (5-1) will describe a straight line which is the control (or transfer) characteristic shown in Fig. 5-4*a*. Actually, β is not a rigorously fixed quantity, so there is an amount of nonlinearity in this characteristic.

There are also *feedback characteristics*, $v_{BE} = f(v_{CE})$ with i_B held constant. They show how changes in the output voltage at a constant input current affect the voltage applied to the transistor input.

Transistors always have internal feedback due to the effect of the bulk base resistance, the base width modulation, and also because the output and input circuits are electrically connected. Therefore, a part of output voltage is always applied to the transistor input. The feedback characteristics are not used for design purposes, and we will leave them out of our discussion. They are even not given in the latest reference sources or data sheets.

Now we will turn to the characteristics of a transistor connected in a common-base circuit.

The input characteristics $i_E = f(v_{EB})$ with v_{CB} held constant (Fig. 5-5) are similar to the characteristic of a forward-biased diode because the emitter current is a forward current. At $v_{CB} = 0$, the curve starts at the origin because there is no current flowing. At $v_{CB} > 0$, the curve is shifted somewhat upwards because the emitter current begins to flow and at $v_{EB} = 0$ there is also the small reverse emitter leakage current, i_{E0}. The equality $v_{EB} = 0$ signifies that the emitter and base are short-circuited. The curves plotted for various values of v_{CB} run very closely to one another, and reference sources usually give only one curve for some normal value of v_{CB}. The slight effect of v_{CB} on the emitter current is explained by the fact that the field set up by v_{CB} is concentrated at the collector junction. Still, an increase in v_{CB} produces a rise in i_E owing to the effect of the bulk base resistance r_{B0}.

It is seen from the circuit diagram of Fig. 5-1*c* that at $E_1 = v_{EB} = 0$, the reverse leakage base current i_{Br} produces across r_{B0} a voltage drop which acts as a forward bias voltage for the emitter junction. Therefore, there appears the reverse emitter leakage current i_{E0} and (as is seen from Fig. 5-1*c*),

$$i_{E0} + i_{Br} = i_{C0}$$

An increase in v_{CB} brings about a rise in i_{Br}, and, as a consequence, there is an increase in the voltage across r_{B0} and in the reverse emitter leakage current i_{E0}. If, on the other hand, the E_1 source feeds a voltage v_{EB} such that i_B is reversed, it will, as usual, be a part of the emitter current (Fig. 5-1*b*). In the circumstances, i_B produces across r_{B0} a voltage drop which opposes the effect of E_1, that is, decreases v_{EB}. An increase in v_{CB} reduces the base width and, as a result, there is a decrease in both i_B and the voltage across r_{B0}. In the final analysis, v_{EB} goes up, and the emitter current also rises.

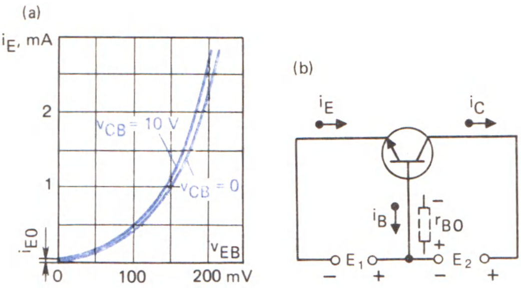


Fig. 5-5

Input characteristics of a transistor connected in a CB circuit

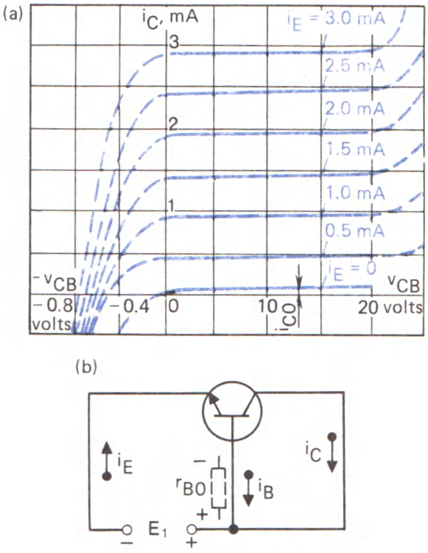


Fig. 5-6

Output characteristics of a transistor connected in a CB circuit

Figure 5-6 shows a family of output characteristics $i_C = f(v_{CB})$ with i_E held constant. They are given for constant values of i_E because

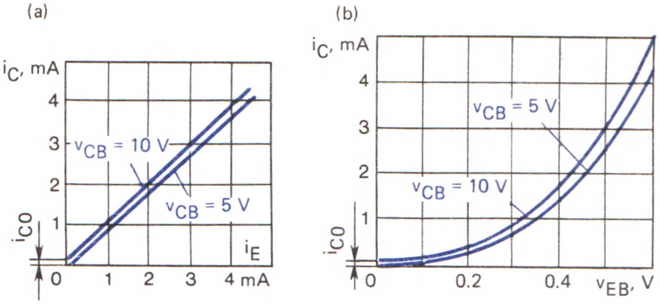


Fig. 5-7

Transfer characteristics of a transistor connected in a CB circuit

the input resistance of a transistor is small and the signal source usually operates as an equivalent constant-current generator, that is, under conditions approaching a short-circuit very nearly. At $i_E = 0$, the curve runs through the origin because, with no emitter current flowing and at $v_{CB} = 0$, no collector current can flow either. This curve is the usual characteristic of a reverse-biased *p-n* junction (diode). The equality $i_E = 0$ signifies that the emitter circuit is open. This in turn implies that only the collector junction biased in the reverse direction is brought in circuit. Under this condition, there is a current flowing which we know as the reverse collector leakage current, i_{CO}.

At some value of v_{CB}, the collector junction breaks down, and there is a sudden surge in the collector current.

Output characteristics plotted for several values of i_E are practically straight lines which run at a very small slope – a fact indicative of a very slight influence of v_{CB} on the collector current. If we wish to raise i_C, we should raise i_E so that more carriers could be injected from emitter to base. But if i_E is held constant, an increase in v_{CB} produces a rise in the collector current mainly owing to a reduction in base width. As a result, fewer injected carriers recombine with the majority carriers of the base region, but a greater number of injected carriers are able to reach the collector so that i_C goes up and i_B falls.

A distinction of output characteristics is the fact that at $v_{CB} = 0$ and $i_E > 0$, the collector current is fairly high, being as large as it is at $v_{CB} > 0$. The explanation is that in this regime there exists some voltage across the collector junction due to the effect of the base resistance r_{B0}. This voltage is produced across r_{B0} by the base current (Fig. 5-6*b*). For many transistors, the output characteristics are straight lines starting at $v_{CB} = 0$. The relation between i_C and i_E is linear very nearly. Therefore, given the same change in i_E, the output curves will be spaced a nearly equal distance apart. The onset of an electrical breakdown is advanced by heavier currents, that is, at lower values of v_{CB}.

The dashed lines in Fig. 5-6*a* show that when the polarity of v_{CB} is reversed, even small values of this voltage will cause the collector current first to fall abruptly and then to reverse its direction of flow and to build up rapidly. This happens because the polarity reversal of v_{CB} as compared with its normal sense renders it a forward-biasing one for the collector junction. When it rises by a few tenths of a volt, it first balances out the small voltage which (as has been explained) exists across the collector junction owing to the voltage drop produced by i_B across the base resistance. Then the voltage across the collector junction is rendered forward-biasing, and i_C builds up at a very high rate in the reverse direction.

Output characteristics for the CB configuration, plotted for several constant values of output voltage v_{EB}, rather than of input current, are not usually employed, and we will not take them up.

The control characteristics for the CB configuration display an almost linear relationship between i_C and i_E (Fig. 5-7*a*). (It should be noted that this relation is more linear than that between i_C and i_B.) These curves plotted for several values of v_{CB} run very closely to one another. This points out that voltage v_{CB} has but a negligible effect on the collector current.

Reference sources usually give only one control characteristic for the average value of v_{CB}. At $i_E = 0$, these curves give the reverse collector leakage current i_{C0}, but ordinarily this current is so small that the curves are shown starting at the origin. In contrast to those shown previously, the curves in Fig. 5-7*b* display a nonlinear relationship between i_C and input voltage. These characteristics are used but seldom. The feedback characteristics $v_{EB} = f(v_{CB})$ with i_E held constant are not practically used, so we will not dwell on them.

The linear relation between i_C and i_E corresponds to the equation derived earlier:

$$i_C = \alpha i_E + i_{C0} \tag{5-2}$$

When α is constant, this equation yields a straight line.

One of the likely test set-ups that can be used to plot the characteristics of *p-n-p* transistors connected in a CE circuit is shown in Fig. 5-8. In this circuit, v_{CE} is adjusted with two variable resistors, R_3 and R_4. The voltage picked off the resistor R_4 is applied to R_3 from which it is fed to the transistor. With such an arrangement, we can obtain a very small voltage v_{CE} and adjust it more gradually. The zero voltage should be set with R_3. The source of E_2 may be a 20-30 V battery or a rectifier. Small values of v_{CE} should be measured, recalling that some voltage will be dropped across the milliammeter intended to measure the collector current.

The base current i_B is measured with a microammeter, and the emitter-to-base voltage v_{BE} with a millivoltmeter. The measurement of voltage at the transistor input may present some difficulties because even high-resistance voltmeters draw a current comparable with the base current. When using the test set-up shown in Fig. 5-8, the true value of v_{BE} can be found by subtracting the voltage drop across the microammeter from the millivoltmeter indication. The voltage drop across the microammeter can readily be determined by multiplying the base current by the resistance of the microammeter. The variable resistor R_2 should have a low resistance (of the order of several tens of ohms). The source of E_1 may be a single dry cell. The purpose of R_1 is to set the voltage across R_2 to a few tenths of a volt. As an alternative, the input circuit may include two variable resistors connected as shown for the collector circuit.

A similar test set-up can be used to measure the characteristics of a transistor connected in a CB circuit. Instead of a microammeter, it should use a milliammeter in order to measure the emitter current.

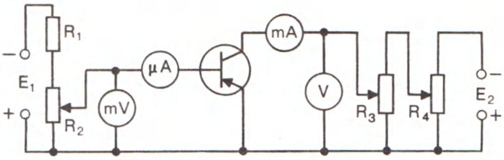


Fig. 5-8

Test set-up to measure transistor characteristics

5-2 Parameters and Equivalent Circuits of Transistors

The parameters, or constants, of a transistor are the quantities that characterize its properties. By using the parameters, we can compare the quality of different transistors, solve problems arising when transistors are used in practical circuits, and design these circuits.

Several sets of parameters and equivalent circuits have been proposed for transistors, each of which has merits and demerits of its own.

All parameters may be classed into *primary* and *secondary*. Primary parameters characterize the properties of a transistor itself, irrespective of the circuit configuration in which it is connected. Secondary parameters vary from one circuit configuration to another.

In addition to the already defined alpha current gain, α, the primary parameters include several resistances in accord with an a.c. equivalent circuit of a transistor (Fig. 5-9). This is what is known as the *r*- or *T*-parameter equivalent circuit. It reflects the electric structure of a transistor and takes into account its amplifying capabilities. In this equivalent circuit (and, indeed, in all other forms of equivalent circuits), it is presumed that a source of the signal to be amplified is connected to the input and supplies an input voltage whose amplitude (peak value) is V_{m1}, while the output is coupled to a load, R_L. Here and elsewhere, we will usually give the peak values of alternating currents and voltages. In many cases they may be replaced with rms or, sometimes, instantaneous values.

The basic primary (*T*- or *r*-) parameters are the resistances r_E, r_C and r_B, called respectively the *emitter resistance*, the *collector resistance*,

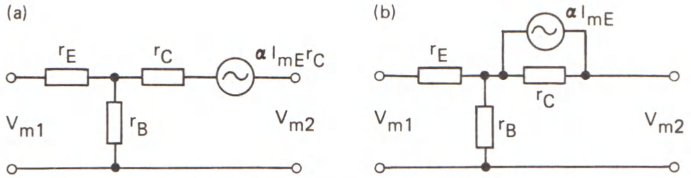


Fig. 5-9

T- (or r-) parameter equivalent circuit of a transistor: (*a*) Thévenin equivalent and (*b*) Norton equivalent

and the *base resistance* to alternating current. The emitter resistance r_E is the resistance presented by the emitter junction to which is added the resistance of the emitter region. Similarly, r_C is the sum of the resistances presented by the collector junction and the collector region, but the latter is negligibly small in comparison with the former. The resistance r_B is the series resistance of the base.

The equivalent circuit we are examining looks like that shown in Fig. 4-4, but the two ought not to be confused. For one thing, the circuit in Fig. 4-4 is not valid for alternating currents because it includes the d. c. resistances r_{E0}, r_{C0} and r_{B0} whereas the a. c. resistances r_E, r_C and r_B are different from their d. c. counterparts owing to the nonlinear properties of transistors. For another, the circuit of Fig. 4-4 does not reflect the amplifying properties of transistors. If we connect a signal source to the input of the circuit in Fig. 4-4, the a. c. voltage emerging across its output will not be boosted – it will be reduced due to the losses across r_{E0} and r_{C0}.

On the contrary, in the circuit of Fig. 5-9*a* the equivalent constant-voltage generator (the Thévenin equivalent) connected in the collector circuit delivers an amplified a. c. voltage; the emf of this generator is proportional to the emitter current i_{mE}.

The equivalent generator must be regarded as an ideal one, and the role of its internal resistance is played by the collector resistance r_C. As will be recalled, the emf of any generator is given by the product between its short-circuit current and internal resistance. In this case, the short-circuit current is αI_{mE} because $\alpha = I_{mC}/I_{mE}$ at $R_L = 0$, that is, when the output is short-circuited. Thus, the emf of the equivalent generator is $\alpha I_{mE} r_C$.

Alternatively, we may use an equivalent constant-current generator (by applying Norton's theorem) instead of an equivalent constant-voltage generator (when we apply Thévenin's theorem). This transformation yields the well-known equivalent circuit shown in Fig. 5-9*b*. In this circuit, the equivalent constant-current generator produces a current equal to αI_{mE}.

Approximately, the values of the primary parameters are as follows. The a. c. emitter resistance r_E is tens of ohms; the a. c. base resistance r_B is several hundred ohms, and the a. c. collector resistance r_C is several hundred kilohms or even several megohms. It is usual to include the alpha current gain α with the other primary parameters.

The equivalent circuit we have just discussed holds for low frequencies only. At high frequencies we must also consider the capacitances of the emitter and collector junctions, and this results in a more elaborate equivalent circuit. This section will only cover low-frequency equivalent circuits and parameters. Operation of transistors at high frequencies will be taken up in Chap. 6.

The equivalent circuit derived by applying Norton's theorem (that is, one containing an equivalent constant-current generator) for a transistor with the common-emitter connection is shown in Fig. 5-10. Here the generator produces a current βI_{mB}, and the resistance of the collector junction is substantially smaller in comparison with what it is in the previous

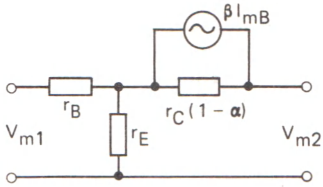


Fig. 5-10

T-parameter equivalent circuit of a transistor connected in the CE configuration equivalent circuit, being $r_C(1-\alpha)$ or, approximately, r_C/β if we recall that $\beta = \alpha/(1-\alpha)$ and $\alpha \approx 1$. The reduction in the resistance of the collector junction in a CE circuit stems from the fact that in this circuit configuration some of v_{CE} is impressed on the emitter junction and boosts carrier injection into the emitter region. In consequence, a noticeable number of injected carriers reach the collector junction and its resistance is brought down.

The equivalent circuit for a CB circuit can be transformed into that for a CE circuit in the following manner. The voltage supplied by any generator is equal to the difference between its emf and the voltage drop across its internal resistance. For the circuit of Fig. 5-9*a* we may write

$$V_m = \alpha I_{mE} r_C - I_{mC} r_C$$

On replacing I$_{mE}$ with a sum, $I_{mC} + i_{mB}$, we obtain

$$\begin{aligned} \mathrm{V}_m &= \alpha(I_{mC} + I_{mB})\, r_C - I_{mC} r_C \\ &= \alpha I_{mC} r_C + \alpha I_{mB} r_C - I_{mC} r_C \\ &= \alpha I_{mB} r_C - (I_{mC} r_C - \alpha I_{mC} r_C) \\ &= \alpha I_{mB} r_C - I_{mC} r_C (1-\alpha) \end{aligned}$$

In the above expression the first term $\alpha I_{mB} r_C$ is the emf, and the second term is the voltage drop due to I_{mC} across $r_C(1-\alpha)$ which is the resistance of the collector junction. The short-circuit current produced by the equivalent constant-current generator is the ratio of the generator emf to its internal resistance, that is,

$$I = \alpha I_{mB} r_C/[r_C(1-\alpha)] = \beta I_{mB}$$

The T-parameter equivalent circuits we have discussed are approximate because the emitter, base and collector are actually connected to one another inside a transistor at more than one point. Still, they are sufficiently accurate when used to solve both theoretical and practical problems.

All sets of secondary parameters are based on the fact that a transistor is treated as a four-terminal (or quadripole or two-port) network, that is, as a device which has two input terminals or poles (the *input port*) and two output terminals or poles (the *output port*). The secondary parameters connect the input and output alternating currents and voltages and hold only for a specified set of operating conditions and low amplitudes. Quite aptly, they are called low-frequency small-signal parameters. Because a transistor is a nonlinear device, changes in its operating conditions and large signals bring about corresponding changes in its secondary parameters.

The secondary parameters most commonly used at this writing are the *h* or *hybrid parameters*. They are called so because their dimensions are mixed, two of them being nondimensional, a third having the dimensions of impedance, and the fourth those of admittance. The *h*-parameters are usually given by manufacturers in specifications and data sheets for transistors. The *h*-parameters are convenient to measure, and this is an important advantage because reference sources usually quote average values derived by measuring the parameters of a large number of transistors of a given type. Two of the *h*-parameters are found with the output short-circuited for a.c., that is, with no load connected to the output. In this case, only a constant voltage (v_2 = const) is passed to the transistor's output from the E_2 source. The remaining two parameters are found with the input circuit open for a.c., that is, when only a constant current (i_1 = const) is flowing in the input circuit, supplied by the respective source. It is an easy matter to hold v_2 and i_1 constant in practice when measuring the *h*-parameters.

More specifically, the *h*-parameters are as follows:

(1) *Input impedance*

$$\mathrm{h}_{11} = \Delta v_1/\Delta i_1 \text{ with } v_2 = \text{const} \qquad (5\text{-}3)$$

is the impedance of a transistor seen at its input terminals by alternating current, with its output short-circuited, that is, with no output alternating voltage present. In the circumstances, the change in input current, Δi_1, is due to a change in only the input voltage, Δv_1. If there were an alternating voltage at the output, it would affect the input current owing to feedback always existing in a transistor. As a result, the input impedance would be different for different values of output alternating voltage which is in turn dependent on the load resistance R_L. However, the h_{11} parameter must characterize a transistor itself (irrespective of R_L), and so it is measured at a constant value of v_2, that is, at $R_L = 0$.

(2) *Reverse voltage feedback ratio*

$$h_{12} = \Delta v_1/\Delta v_2 \text{ with } i_1 = \text{const} \qquad (5\text{-}4)$$

shows what fraction of output alternating voltage is fed back to a transistor's input owing to the internal feedback always existing there. The condition $i_1 = \text{const}$ in this case stresses the fact that there is no alternating current flowing in the input circuit. In other words, the input circuit is open for a. c. and, in consequence, Δv_1 is solely due to Δv_2.

We have noted more than once the existence of internal feedback in transistors because the transistor electrodes are electrically connected to one another, and also because the base has an inherent resistance. This feedback exists at any frequency, however low, even at $f = 0$.

(3) *Forward current transfer ratio*

$$h_{21} = \Delta i_2/\Delta i_1 \text{ with } v_2 = \text{const} \tag{5-5}$$

gives how much a transistor amplifies alternating current in operation at no-load. The condition $v_2 = \text{const}$, that is, $R_L = 0$, is specified here for the change in output current, Δi_2, to be solely a function of a change in input current, Δi_1. It is only then that the h_{21} parameter does characterize current amplification by a transistor itself. If the output voltage were varying, it would affect the output current, and we would not be able to evaluate gain from changes in this current.

(4) *Output admittance*

$$h_{22} = \Delta i_2/\Delta v_2 \text{ with } i_1 = \text{const} \tag{5-6}$$

is the internal admittance for a. c. between the output terminals of a transistor with its input open-circuited. Output current i_2 should change solely due to changes in output voltage, v_2. If input current i_1 were not constant, its changes would cause changes in i_2, and a wrong value would be found for h_{22}.

The h_{22} parameter is expressed in siemenses (S). Since admittance will in our subsequent calculations be used more seldom than impedance, we will often replace the h_{22} parameter with its reciprocal, or the output impedance (assumed to be purely resistive), $R_{out} = 1/h_{22}$, expressed in ohms or kilohms.

So far we have defined the h-parameters in terms of incremental changes in currents and voltages. An alternative way is to define them in terms of the peak values (amplitudes) of alternating currents and voltages:

$$h_{11} = V_{m1}/I_{m1} \text{ with } V_{m2} = 0 \tag{5-7}$$

$$h_{12} = V_{m1}/V_{m2} \text{ with } I_{m1} = 0 \tag{5-8}$$

$$h_{21} = I_{m2}/I_{m1} \text{ with } V_{m2} = 0 \tag{5-9}$$

$$h_{22} = I_{m2}/V_{m2} \text{ with } I_{m1} = 0 \tag{5-10}$$

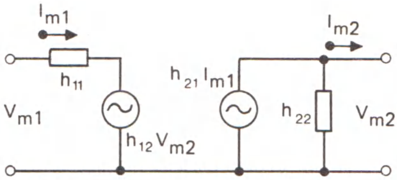


Fig. 5-11

Hybrid-parameter equivalent circuit of a transistor

It is to be stressed once again that the h-parameters are only valid for small-signal conditions. Their use under large-signal conditions would yield results which are greatly in error.

When measuring the h-parameters at alternating current, we may replace the peak values with the respective rms values as indicated by meters.

The relations between alternating currents and voltages in a transistor may be expressed in terms of the h-parameters as follows:

$$V_{m1} = h_{11}I_{m1} + h_{12}V_{m2} \tag{5-11}$$

$$I_{m2} = h_{21}I_{m1} + h_{22}V_{m2} \tag{5-12}$$

This is because the input voltage V_{m1} is the sum of the voltage drop produced by the input current I_{m1} across the input impedance h_{11}, and the voltage which is passed on from output to input due to feedback and is a component of the output voltage V_{m2}. This part is represented by the h_{12} parameter. The output current I_{m2} is the sum of the amplified current $h_{21}I_{m1}$ and the current produced in the h_{22} circuit element by the output voltage V_{m2}.

Equations (5-11) and (5-12) apply to the equivalent circuit shown in Fig. 5-11. Here the constant-voltage generator $h_{12}V_{m2}$ represents the feedback voltage existing in the input circuit. The generator itself should be assumed to be ideal, that is, with an internal resistance equal to zero. The equivalent constant-current generator $h_{21}I_{m1}$ in the output circuit accounts for the effect of current amplification, and the h_{22} parameter is the internal admittance (or, rather, conductance) of the current generator. Although the input and output circuits do not appear interconnected, actually the two equivalent generators do take care of the interrelation that exists between the two circuits.

The h-parameters are further supplied with letter subscripts in order to identify the circuit configuration to which they apply, namely the letter '*e*' for the CE connection, the letter '*b*' for the CB connection, and the letter 'c' for the CC connection.

Let us discuss the *h*-parameters for the CE and CB circuits and see what values they have for low-power transistors.

For a CE circuit, $i_1 = i_B$, $i_2 = i_C$, $v_1 = v_{BE}$, $v_2 = v_{CE}$, so the *h*-parameters can be defined as follows: – Input impedance

$$h_{11e} = \Delta v_{BE}/\Delta i_B \text{ with } v_{CE} = \text{const} \qquad (5\text{-}13)$$

It ranges in value from several hundred ohms to several kilohms.

– Reverse voltage feedback ratio

$$h_{12e} = \Delta v_{BE}/\Delta v_{CE} \text{ with } i_B = \text{const} \qquad (5\text{-}14)$$

It usually is 10^{-3}-10^{-4} which means that the voltage fed back from output to input is a few thousandths or ten-thousandths of the output voltage.

– Forward current transfer ratio

$$h_{21e} = \beta = \Delta i_C/\Delta i_B \text{ with } v_{CE} = \text{const} \qquad (5\text{-}15)$$

It is anywhere from a few tens to several hundreds.

– Output admittance

$$h_{22e} = \Delta i_C/\Delta v_{CE} \text{ with } i_B = \text{const} \qquad (5\text{-}16)$$

It is equal to a few tenths or hundredths of a millisiemens so that the output impedance (resistance), $1/h_{22e}$, is from several kilohms to a few tens of kilohms.

For a CB circuit,

$$i_1 = i_E,\ i_2 = i_C,\ v_1 = v_{EB}, \text{ and } v_2 = v_{CB}$$

Therefore, the respective *h*-parameters will be written as follows:

– Input impedance

$$h_{11b} = \Delta v_{EB}/\Delta i_E \text{ with } v_{CB} = \text{const} \qquad (5\text{-}17)$$

It has a value of units or tens of ohms.

– Reverse voltage feedback ratio

$$h_{12b} = \Delta v_{EB}/\Delta v_{CB} \text{ with } i_E = \text{const} \qquad (5\text{-}18)$$

It has the same order of magnitude (10^{-3}-10^{-4}) as in the CE circuit.

– Forward current transfer ratio

$$|h_{21b}| = \alpha = \Delta i_C/\Delta i_E \text{ with } v_{CB} = \text{const} \qquad (5\text{-}19)$$

It usually is 0.950-0.998.*

– Output admittance

$$h_{22b} = \Delta i_C/\Delta v_{CB} \text{ with } i_E = \text{const} \qquad (5\text{-}20)$$

It has a value of several microsiemenses or even less; the output impedance (resistance), $1/h_{22b}$, is usually as high as several hundred kilohms which is noticeably greater than it is in the CE connection.

With any circuit configuration, the *h*-parameters can be converted to the primary (*T*- or *r*-) parameters of a transistor in a well-defined manner. Thus, for a CB circuit,

$$\begin{aligned} h_{11b} &\approx r_E + r_B(1-\alpha) \\ h_{21b} &= -\alpha \\ h_{12b} &\approx r_B/r_C \\ h_{22b} &\approx 1/r_C \end{aligned} \qquad (5\text{-}21)$$

For a CE circuit,

$$\begin{aligned} h_{11e} &= r_B + r_E/(1-\alpha) \\ h_{12e} &\approx r_E/r_C(1-\alpha) \\ h_{21e} &= \beta = \alpha/(1-\alpha) \\ h_{22e} &= 1/r_C(1-\alpha) \end{aligned} \qquad (5\text{-}22)$$

If the *h*-parameters are known, it is an easy matter to find the primary (*T*- or *r*-) parameters of a transistor.

The equations connecting the various parameters are derived from inspection of the respective equivalent circuits. Taking the equivalent circuit of Fig. 5-9 as an example, we may write

$$h_{11b} = \left.\frac{V_{mEB}}{I_{mE}}\right|_{V_{mCE}=0} = \frac{I_{mE}r_E + I_{mB}r_B}{I_{mE}}$$

$$= r_E + \frac{I_{mE} - I_{mC}}{I_{mE}} r_B = r_E + (1-\alpha)\, r_B$$

$$h_{12b} = \left.\frac{V_{mEB}}{V_{mCB}}\right|_{I_{mE}=0} = r_B/(r_B + r_C) \approx r_B/r_C$$

because r_B is a very small fraction of r_C;

$$h_{21b} = \left.\frac{I_{mC}}{I_{mE}}\right|_{V_{mCB}=0} = -\alpha$$

$$h_{22b} = \left.\frac{I_{mC}}{V_{mCB}}\right|_{I_{mE}=0} = 1/(r_C + r_B) \approx 1/r_C$$

Similarly, we can derive equations for the CE circuit of Fig. 5-10.

The values of the *h*-parameters for the CE and

* The currents i_E and i_C take opposite signs because one enterns and other leaves the transistor, and so h_{21b} takes a "$-$" sign, that is, $h_{21b} = -\alpha$.

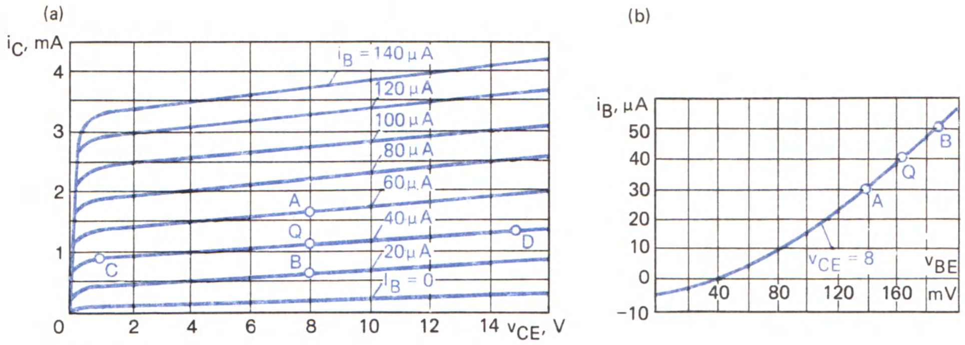


Fig. 5-12

Transistor connected in a CE circuit: (*a*) output characteristic and (*b*) input characteristic

CB circuits are summarized in Table 5-1 where h_{22} has been replaced with $1/h_{22}$.

If we are given the output characteristics of a transistor (Fig. 5-12*a*), we can determine the h_{21e} and h_{22e} parameters for some operating point Q. The procedure is as follows. Taking the incremental changes Δi_C and Δi_B between points A and B and assuming a constant value for v_{CE}, we may write

$$h_{21e} = \beta = \Delta i_C/\Delta i_B = 1 \text{ mA}/40 \text{ μA} = 25$$

The ratio of the incremental changes Δi_C and Δv_{CE} between points C and D, assuming a constant value for i_B, gives

$$h_{22e} = \Delta i_C/\Delta v_{CE}$$
$$= 0.4 \times 10^{-3}/14 = 28.6 \times 10^{-6} \text{ S}$$

which corresponds to the output impedance given by

$$1/h_{22e} = 1/28.6 \times 10^{-6} \text{ S} \approx 36\,200\ \Omega \approx 36 \text{ k}\Omega$$

Figure 5-12*b* shows the input characteristic of the same transistor on which the Q-point is specified for the same conditions as it is on the output characteristics. Taking the increments Δv_{BE} and Δi_B between points A and B and assuming a constant value for v_{CE}, we obtain

$$h_{11e} = \Delta v_{BE}/\Delta i_B = 50 \text{ mV}/20 \text{ μA} = 2.5 \text{ k}\Omega$$

Table 5-1

VALUES OF *h*-PARAMETERS

Parameter	CE circuit	CB circuit
h_{11}	Hundreds of ohms to units of kilohms	Units to tens of ohms
h_{12}	10^{-3}-10^{-4}	10^{-3}-10^{-4}
$\lvert h_{21} \rvert$	Tens to hundreds (β)	0.950-0.998 (α)
$1/h_{22}$	Units to tens of kilohms	Hundreds of kilohms to several megohms

In order to find h_{12e}, we need at least two input characteristics plotted for different values of v_{CE}. Data sheets, however, usually give only one characteristic from which h_{12e} cannot be found (one ought not to use the input characteristic plotted for $v_{CE} = 0$ in order to find the h-parameters). Therefore, the h_{12e} parameter is not used in simple practical calculations, so we will not be concerned with its determination from a characteristic.

In the literature on the subject the number subscripts of the h-parameters are often replaced with letter subscripts as follows:

(1) The first subscript indicates the characteristic: *i* for input, *o* for output, *f* for forward transfer, and *r* for reverse transfer.

(2) The second subscript indicates the circuit configuration: *b* for common-base (CB), *c* for common-collector (CC), and *e* for common-emitter (CE).

Using this system, h_{11} or input resistance can be indicated by h_{ib} for common-base input, h_{ie} for common-emitter, and h_{ic} for common-collector. Similarly, h_{12} or the reverse transfer voltage ratio can be indicated by h_{rb}, h_{re} and h_{rc} for common-base, common-emitter and common-collector, respectively. By the same token, h_{21}, or the forward current transfer ratio, can be indicated by h_{fb}, h_{fe}, and h_{fc}. Finally, h_{22} or the output admittance (most often, conductance), can be indicated by h_{ob}, h_{oe}, and h_{oc}, respectively, for common-base, common-emitter and common-collector.

When the h-parameters were first used, the CB configuration was the most popular. Today, the CE configuration is used to the greatest extent. For this reason, practically all two-junction transistor data sheets outside the USSR list h_{fe}. Some data sheets mix the configurations. For example, a typical data sheet for two-junction transistors will quote h_{fe}, h_{ib}, h_{rb}, and h_{ob}.

Another system of parameters uses what are known as *admittance* or *y-parameters*. In general they are used with FETs. This is practically true when FETs are used at high (radio) frequencies. At low frequencies, admittances may safely be taken as being pure conductances and the y-parameters may be replaced by what are known as the *conductance* or *g-parameters*, each supplied with applicable subscripts as already explained. These parameters are measured with the input or the output short-circuited for a. c., using the following equations.

(1) *Input admittance*

$$y_{11} = \Delta i_1/\Delta v_1 \text{ with } v_2 = \text{const} \qquad (5\text{-}23)$$

It is easy to see that y_{11} is the inverse of h_{11}:

$$y_{11} = 1/h_{11} \qquad (5\text{-}24)$$

(2) *Reverse transfer admittance*

$$y_{12} = \Delta i_1/\Delta v_2 \text{ with } v_1 = \text{const} \qquad (5\text{-}25)$$

The y_{12} parameter shows what change in i_1 results due to feedback when v_2 changes by 1 V.

(3) *Forward transfer admittance* (or *transconductance*)

$$y_{21} = \Delta i_2/\Delta v_1 \text{ with } v_2 = \text{const} \qquad (5\text{-}26)$$

The y_{21} parameter describes the control action of input voltage v_1 on output current i_2 and defines the change in i_2 produced by a change of 1 V in v_1. It ranges in value from tens to hundreds of milliamperes per volt (millisiemens).

(4) *Output admittance*

$$y_{22} = \Delta i_2/\Delta v_2 \text{ with } v_1 = \text{const} \qquad (5\text{-}27)$$

It is to be noted that y_{22} and h_{22} are different quantities because they are found under different conditions (at $v_1 = \text{const}$ and $i_1 = \text{const}$, respectively).

The y_{21} parameter is connected to the h-parameters by a simple relation

$$y_{21} = h_{21}/h_{11} \qquad (5\text{-}28)$$

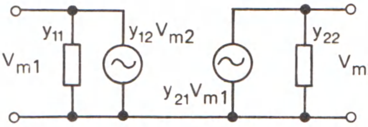


Fig. 5-13

Admittance (y)-parameter equivalent circuit of a transistor

We leave it as an exercise for the reader to check the above equation.

Sometimes, the system of y-parameters is extended to include the static voltage gain of a transistor:

$$\mu = -\Delta v_2/\Delta v_1 \text{ with } i_2 = \text{const*} \qquad (5\text{-}29)$$

The μ-parameter is related to the other y-parameters by a relation of the form

$$\mu = y_{21}/y_{22} \qquad (5\text{-}30)$$

and is equal to thousands for transistors.

Using the y-parameters, we can connect the currents and voltages existing in a transistor by the following equations:

$$I_{m1} = y_{11}V_{m1} + y_{12}V_{m2} \qquad (5\text{-}31)$$

$$I_{m2} = y_{21}V_{m1} + y_{22}V_{m2} \qquad (5\text{-}32)$$

Equations (5-31) and (5-32) show that input current I_{m1} is the sum of the current produced by input voltage V_{m1} across the y_{11} element in the circuit, and the current which arises in the input circuit due to V_{m2} and feedback. Output current I_{m2} is the sum of the amplified current $y_{21}V_{m1}$ and the current produced in the y_{22} element by V_{m2}.

The equivalent circuit that applies to the y-parameters is shown in Fig. 5-13 and is described by Eqs. (5-31) and (5-32). In this circuit, the equivalent constant-current generator $y_{21}V_{m1}$ accounts for the amplification produced by the transistor, and the equivalent constant-current generator $y_{12}V_{m2}$ takes care of the internal feedback existing in transistors. Sometimes a transistor may be represented by an equivalent pi-circuit in which the admittances (Fig. 5-14) are connected to the y-pa-

* The collector current will remain constant only if changes in the voltages are opposite in sign. Therefore, there is a "−" sign in Eq. (5-29). The gain factor μ itself is a positive quantity.

rameters in the following manner:

$$
\begin{aligned}
y_1 &= y_{11} + y_{12} \\
y_2 &= y_{22} + y_{12} \\
y_0 &= -y_{12} \qquad (5\text{-}33) \\
y &= y_{21} - y_{12}
\end{aligned}
$$

The constant-current generator yV_{m1} in this equivalent circuit accounts for the amplified current produced in the output circuit.

An advantage of the y-parameters is that they are similar to the parameters of vacuum tubes. Their disadvantage consists in that for measuring y_{12} and y_{22} the input must be short-circuited for a. c., and this is very difficult to do in practice because the input resistance of the transistor itself is low and that of the instrument used to measure the input current (a microammeter or a milliammeter) is not able to short-circuit the input.

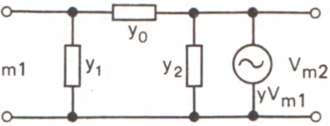


Fig. 5-14
Equivalent pi-circuit of a transistor

Chapter Six
Dynamic Operation of Bipolar Transistors

6-1 Calculation of Dynamic Operation of Transistors

Dynamic operation of transistors, that is, when they are amplifying an a. c. signal, has in part been discussed in Chap. 4. In this mode, a load, R_L, is connected to the output circuit. As a rule, the load resistance is a small fraction of the output resistance R_{out} of the transistor itself. Notably, this is the case when the load is shunted by the low input resistance of the next stage. In such cases, calculations are simplified by arbitrarily assuming that the transistor is operating at no-load.

Amplification always entails a varying degree of distortion, depending on the conditions under which the signal source is operating. We will consider two cases which are most typical. Let the signal source generate a sinewave emf

$$e_{in} = E_{m\,in} \sin \omega t$$

and have an internal resistance R_{ss}. Let us also agree that this resistance and the load resistance R_L are linear. The input resistance of the transistor, R_{in}, is, as will be recalled, nonlinear because the input characteristic $i_{in} = f(v_{in})$, which reflects the nonlinear properties of the transistor itself, is likewise nonlinear.

Since R_{in} is small, it happens most often that R_{in} is a small fraction of R_{ss}; in consequence, the signal source is operating as an equivalent constant-current generator under conditions very close to a short circuit. The input alternating current in this case is

$$i_{in} \approx e_{in}/R_{ss}$$

and has a sinewave form because the input emf e_{in} is sinusoidal and R_{ss} is linear. The output alternating current is approximately proportional to the input current and is likewise sinusoidal. It is obvious that the output voltage

$$v_{out} = i_{out} R_L$$

must be sinusoidal. In the circumstances, the input signal is only slightly corrupted by nonlinear distortion while it is amplified. Importantly, although the input voltage

$$v_{in} = i_{in} R_{in}$$

is distorted (is other than sinewave in shape) because R_{in} is nonlinear, the amplified signal emerging at the input is nearly nondistorted. The insignificant nonlinear distortion that does occur is due to the fact that i_{out} is not an exactly linear function of i_{in}.

A more infrequent occurrence is when R_{in} is many times R_{ss} because signal sources having a low internal resistance are rare. In the circumstances,

$$i_{\text{in}} \approx e_{\text{in}}/R_{\text{in}}$$

and its shape is other than sinusoidal because R_{in} is nonlinear. But the output current proportional to the input current will then be likewise nonsinusoidal, and the output signal will therefore be distorted despite the fact that the input voltage in this mode of operation is nearly equal to the emf and is sinewave in shape.

Simple calculations yield only approximate values of operating conditions. This is acceptable in many cases because there is always a noticeable spread in ratings among transistors.

If R_{L} is a small fraction of R_{out}, the current gain k_I is approximately equal to h_{21}, so that k_I is nearly equal to the alpha current gain for a CB circuit and to the beta current gain for a CE circuit.

The stage voltage gain is given by

$$k_V = V_{m\,\text{out}}/V_{m\,\text{in}} = I_{m\,\text{out}}R_{\text{L}}/I_{m\,\text{in}}R_{\text{in}}$$
$$= k_I R_{\text{L}}/R_{\text{in}} \qquad (6\text{-}1)$$

The input resistance of a stage may approximately be taken equal to the h_{11} parameter of the transistor:

$$R_{\text{in}} \approx h_{11} \qquad (6\text{-}2)$$

Then,

$$k_V \approx R_{\text{L}}h_{21}/h_{11} \qquad (6\text{-}3)$$

However,

$$h_{21}/h_{11} = y_{21}$$

Therefore,

$$k_V \approx y_{21}R_{\text{L}} \qquad (6\text{-}4)$$

For a more accurate calculation of the dynamic operation (amplification mode), we may use the following equations:

$$V_{m1} = h_{11}I_{m1} + h_{12}V_{m2} \qquad (6\text{-}5)$$

$$I_{m2} = h_{21}I_{m1} + h_{22}V_{m2} \qquad (6\text{-}6)$$

We can express V_{m2} in terms of I_{m2} on recalling that

$$v_2 = E_2 - i_2R_{\text{L}}$$

Then,

$$\Delta v_2 = -\Delta i_2 R_{\text{L}}$$

because the incremental change in the constant quantity E_2 is zero. Incremental changes may be treated as peak values or amplitudes. Hence,

$$V_{m2} = -I_{m2}R_{\text{L}}$$

The "$-$" sign indicates that there is a phase shift of 180° between the incremental changes in v_2 and i_2. On replacing V_{m2} with $-I_{m2}R_{\text{L}}$, we may re-write Eqs. (6-5) and (6-6) as

$$V_{m1} = h_{11}I_{m1} - h_{12}I_{m2}R_{\text{L}} \qquad (6\text{-}7)$$

$$I_{m2} = h_{21}I_{m1} - h_{22}I_{m2}R_{\text{L}} \qquad (6\text{-}8)$$

Solving Eq. (6-8) for I_{m2} gives

$$I_{m2} + h_{22}I_{m2}R_{\text{L}} = h_{21}I_{m1}$$

$$I_{m2}(1 + h_{22}R_{\text{L}}) = h_{21}I_{m1}$$

On dividing both sides of the equality by $(1 + h_{22}R_{\text{L}})$ and by I_{m1}, we get

$$I_{m2}/I_{m1} = k_I = h_{21}/(1 + h_{22}R_{\text{L}})$$
$$= h_{21}R_{\text{out}}/(R_{\text{out}} + R_{\text{L}}) \qquad (6\text{-}9)$$

At $R_{\text{L}} \ll R_{\text{out}}$, we find that $\text{k}_I \approx h_{21}$.

On dividing both sides of Eq. (6-7) by I_{m1}, we obtain an equation for R_{in}

$$V_{m1}/I_{m1} = R_{\text{in}} = h_{11} - h_{12}k_IR_{\text{L}} \qquad (6\text{-}10)$$

When R_{L} is small and recalling that h_{12} is low in value (a small fraction of unity), we will have $R_{\text{in}} \approx h_{11}$ in many cases.

Knowing k_I and k_V and given the input current or input voltage, we are able to calculate the output current and voltage, the input and output power, and the power gain. For example, given $I_{m\,\text{in}}$, we get

$$V_{m\,\text{in}} = I_{m\,\text{in}}R_{\text{in}} \approx I_{m\,\text{in}}h_{11} \qquad (6\text{-}11)$$

$$P_{\text{in}} = I_{m\,\text{in}}V_{m\,\text{in}}/2 \qquad (6\text{-}12)$$

$$\text{I}_{m\,\text{out}} = k_I I_{m\,\text{in}} \approx h_{21}I_{m\,\text{in}} \qquad (6\text{-}13)$$

$$V_{m\,\text{out}} = k_V V_{m\,\text{in}} \text{ or } V_{m\,\text{out}} = I_{m\,\text{out}}R_{\text{L}} \qquad (6\text{-}14)$$

$$P_{\text{out}} = I_{m\,\text{out}}V_{m\,\text{out}}/2 \qquad (6\text{-}15)$$

$$k_P = k_Ik_V \text{ or } k_P = P_{\text{out}}/P_{\text{in}} \qquad (6\text{-}16)$$

The above simple procedure for the calculation of dynamic operation on the basis of transistor parameters applies under small-signal conditions when incremental changes in signal amplitudes cannot be plotted as curves and a grapho-analytical solution is out of the question.

Sometimes, the gain of a transistor stage is defined as the ratio of output voltage to the emf (terminal voltage) of the signal source, $E_{m\,\text{in}}$. There is a good deal of sense in this approach because $V_{m\,\text{in}}$ is, as a rule, substantially smaller than $E_{m\,\text{in}}$ because of the low input resistance of the transistor. The calculation of the stage power gain is then changed accordingly. The

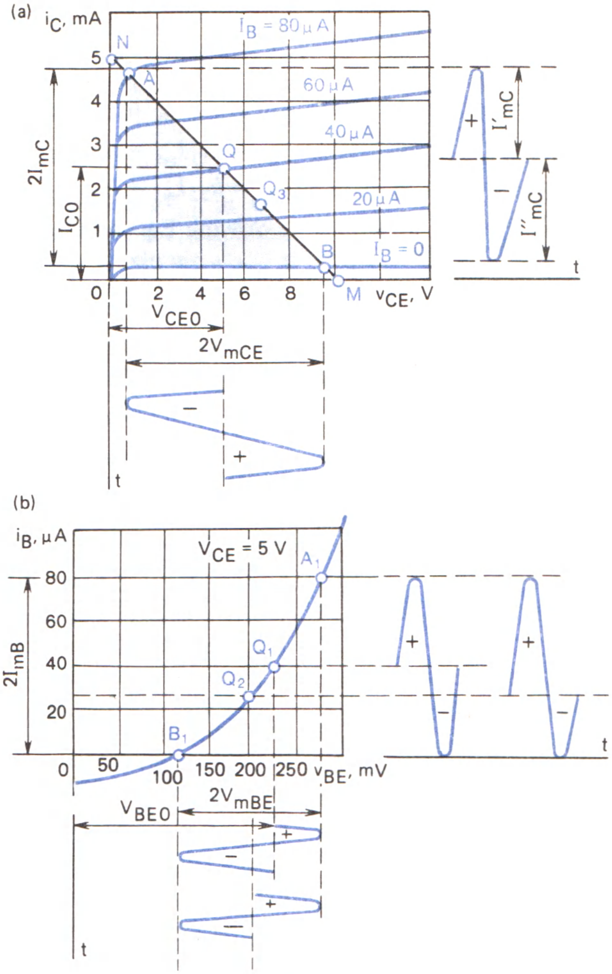


Fig. 6-1

Grapho-analytic study of a transistor's operation at load, using its output and input characteristics

values of k_V and k_P thus found will depend on the relative values of the transistor's input resistance and the signal source resistance R_{ss}.

Now we will consider a grapho-analytical procedure used to determine the dynamic operation of a transistor. This procedure is more accurate because it takes into account the nonlinear properties of transistors. Also, the grapho-analytical procedure yields a more detailed set of design data: it determines the quantities connected to both the alternating and the direct components of currents and voltages.

The grapho-analytical procedure is based on *dynamic* or *load characteristics*. Because a transistor is a current-controlled device, we need to use both input and output characteristics. As an example, we will take these characteristics for a CE stage in which the load presents the same resistance, R_L, to both d.c. and a.c.

On a family of output characteristics (Fig. 6-1*a*), the dynamic characteristic (otherwise called the *load line*) is constructed using specified or chosen values of the E_2 supply voltage and load resistance R_L.

Because the output circuit satisfies the equation

$$E_2 = v_{CE} + i_C R_L \qquad (6\text{-}17)$$

the load line is constructed, using its intersections with the coordinate axes, that is, in the same way as we did it for the diode (see Sec. 3-4). At $i_C = 0$, we obtain $E_2 = v_{CE}$, that is, we lay off E_2 on the voltage axis (this yields point M). At $v_{CE} = 0$, we get $i_C = E_2/R_L$ and lay off this value on the current axis (this yields point N). On joining the two points M and N, we obtain the desired load line. The next step is to choose the operating region on the load line. For example, if we wish to derive a large output power, the operating region of choice should be portion AB. The intercepts of the operating region on the coordinate axes define the double amplitudes of the fundamental components of alternating output current and voltage, $2I_{mC}$ and $2V_{mCE}$. Now, it is an easy matter to determine output power

$$P_{out} = I_{mC}V_{mCE}/2 \tag{6-18}$$

The shaded area in Fig. 6-1*a* is what may be called the available power triangle. Its hypotenuse is the operating region AB, and its sides are, respectively, the double amplitude of current, $2I_{mC}$, and of voltage $2V_{mCE}$. It is an easy matter to calculate that the area of the triangle is four times the available power, $2I_{mC}V_{mCE}$.

Let the signal source resistance R_{ss} be many times the transistor input resistance R_{in}. Then the nonlinearity of R_{in} may practically be neglected because the properties of the input circuit will be determined by R_{ss}. If it is linear, then with a sinewave signal emf, the input current will likewise be sinewave. Therefore the operating point Q corresponds to a current, I_{B0}, which is the arithmetic mean of the currents at points A and B. The Q- (or quiescent) point defines the fundamental amplitude of input current, I_{mB}, as half the difference between the base currents corresponding to points A and B, and also I_{C0} and V_{CE0} in the *quiescent state*. From these values we can find P_{C0} dissipated in the transistor in the quiescent state – it ought not to exceed the absolute maximum power rating, $P_{C\,max}$, which is one of the key ratings of transistors:

$$P_{C0} = I_{C0}\,V_{CE0} \leqslant P_{C\,max} \tag{6-19}$$

If the designer has at his disposal a family of input characteristics for the transistor he has chosen, he can plot the input load line by transferring the output load line point-by-point onto that family of curves. It is usual, however, for reference sources (such as data sheets) to give only curves for $v_{CE} = 0$ and for some $v_{CE} > 0$ or only the last one. Because input characteristics for stepped values of v_{CE} in excess of 0.5-1 V are spaced closely apart, the load line differs only slightly from them. Therefore input currents and voltages may approximately be calculated from the input characteristic for $v_{CE} > 0$ taken from a reference source or a data sheet. Points A, Q, and B of the output load line are translated onto that curve to locate points A_1, Q_1 and B_1 (Fig. 6-1*b*). The projection of the operating region A_1B_1 on the voltage axis defines the double amplitude of input voltage, $2V_{mBE}$. Knowing I_{mB} and V_{mBE}, we can readily calculate the input resistance R_{in} and the input power P_{in} of the stage by the equations

$$R_{in} = V_{mBE}/I_{mB} \tag{6-20}$$

$$P_{in} = V_{mBE}I_{mB}/2 \tag{6-21}$$

The operating point Q_1 also defines the d.c. base voltage V_{BE0}. Knowing V_{BE0} and assuming approximately that the direct component of base current in the dynamic operation is I_{B0}, it is an easy matter to calculate the resistance of the swamping resistor R_B via which a direct voltage will be applied from the E_2 source to the base:

$$R_B = (E_2 - V_{BE0})/I_{B0} \tag{6-22}$$

The stage current, voltage, and power gains can then be found by the usual equations

$$\begin{aligned} k_I &= I_{mC}/I_{mB} \\ k_V &= V_{mCE}/V_{mBE} \\ k_P &= k_I k_V \end{aligned} \tag{6-23}$$

It may be taken approximately that the direct component of collector current in the dynamic operation is equal to the quiescent current I_{C0}. Then the power expended by the E_2 source can be found by the equation *

$$P_0 = E_2I_{C0} \tag{6-24}$$

and the stage efficiency (or, to be more accurate, the efficiency of the output circuit) will be given by

$$\eta = P_{out}/P_0 \tag{6-25}$$

It is shown in Fig. 6-1*b* that at the operating point Q_1 the input current is distorted very little: both of its half-cycles have the same amplitude. In contrast, the input voltage is heavily distorted: its positive half-cycle is substantially smaller in amplitude than its negative half-cycle. Still, the output current and the output voltage

* The power supplied to the base circuit may be neglected as being extremely small.

are only slightly distorted. As has been shown, this result is typical when the signal source is acting as an equivalent constant-current generator (at $R_{ss} \gg R_{in}$) and is feeding a sinewave current to the transistor input. If, on the other hand, the signal source is operating as an equivalent constant-voltage generator (at $R_{ss} \ll R_{in}$) and is feeding a sinusoidal voltage to the input, the operating point will be located at Q_2, and the input current will be heavily distorted. The output current and the output voltage will likewise be heavily distorted, because the operating point will be located at Q_3 on the output characteristics, thereby dividing the operating region AB into two unequal parts.

When the positive and negative half-cycles of collector current differ in amplitude (let them be denoted as I'_{mC} and I''_{mC}), we can find the second harmonic amplitude of this current, I_{mC2}, and the incremental change in its direct component, ΔI_{C0}, by the equation

$$I_{mC2} = \Delta I_{C0} = (I'_{mC} - I''_{mC})/4 \qquad (6\text{-}26)$$

Then the direct component (average value) of collector current in the dynamic operation will be

$$I_{C\ av} = I_{C0} + \Delta I_{C0} \qquad (6\text{-}27)$$

For the CE circuit it is usual that $I'_{mC} < I''_{mC}$. Therefore, $\Delta I_{C0} < 0$ and $I_{C\ av} < I_{C0}$.

The change occurring in the direct component of collector current as the transistor moves from the quiescent state to the dynamic operation is an indication of nonlinear distortion. If the milliammeter measuring this current gives the same reading with the signal applied to and removed from the input, distortion is practically nonexistent.

The principles of graphical analysis examined for low-distortion amplification apply to other locations of the operating point as well. They are somewhat different for a transformer-coupled load and a load in the form of a resonant (tank) circuit (see Figs. 4-14 and 4-15). In such cases, the load line is constructed in a different way. This is because a resonant circuit or a loaded transformer present different impedances to the direct and alternating components of collector current. The tank-circuit coil or the transformer primary presents a relatively low resistance to direct current. We may neglect the loss of some d. c. supply voltage across this resistance and deem approximately that the direct collector voltage V_{CE0} is equal to the supply voltage:

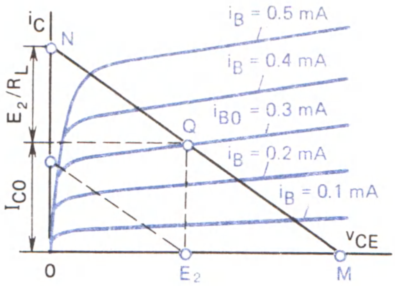


Fig. 6-2

Construction of the load line for a transformer-coupled or an inductively coupled amplifier stage

$$V_{CE0} \approx E_2 \qquad (6\text{-}28)$$

The situation is different with the alternating component of collector current to which a resonant (tank) circuit presents a very high impedance – of the order of several thousand or even tens of thousands of ohms. The primary of a loaded transformer presents about the same impedance to alternating current. In consequence, the transistor is operating at no-load in terms of direct current, and at load in terms of alternating current. The basic equation defining the operation of a transistor at load, Eq. (6-17), must now be written differently

$$v_{CE} = E_2 - \Delta i_C R_L \qquad (6\text{-}29)$$

Instead of collector current, we must consider its incremental change, Δi_C, which is in effect an alternating component, because the load resistance R_L only exists for this component. Here Δi_C should be construed as a change of current occurring at a sufficiently high frequency, say at the resonance frequency of the tank circuit, because it is only then that a resonant circuit possesses a high and purely resistive impedance.

In order to plot the applicable load line, we set in Eq. (6-29) $\Delta i_C = 0$ so that $v_{CE} = E_2$. This case corresponds to the location of the operating point at Q (Fig. 6-2) defining the quiescent state. Before it can be plotted on the diagram, however, we need to know the direct component of base current, I_{B0}. Point Q defines the quiescent current I_{C0}. The other point of the load line can be located on setting $v_{CE} = 0$. Then,

$$\Delta i_C = E_2/R_L$$

and the intercept on the axis of ordinates will locate point N which is needed only for purposes of construction. This point (and a number of points around it) does not correspond to any practical operating conditions because at $v_{CE} = 0$ the collector current in a transistor cannot be a maximum. Now draw a straight line through points Q and N – it is the load line. For comparison, the diagram also includes a dashed line – it is the load line of a resistance-coupled stage with the same value of R_L, that is, when R_L is the same for direct and alternating currents. This load line is shifted downwards by an amount equal to the quiescent collector current, I_{C0}.

The load lines for transformer-coupled and tank-circuit stages differ in the following aspects. The operating point corresponds to E_2 rather than to $V_{CE0} = E_2 - I_{C0}R_L$. When constructing the load line for a resistance-coupled stage, we laid off an intercept E_2/R_L on the current axis from the origin; now we lay it off from I_{C0} so that the line runs somewhat higher. It is interesting to note that during the negative half-cycles of input current, when the collector current decreases ($\Delta i_C < 0$ and $i_C < I_{C0}$), the collector voltage exceeds E_2. Over the entire region QM of the load line, the collector voltage exceeds the supply voltage.

This situation, which might seem strange at first glance, is explained by the fact that the collector circuit contains energy-storing elements – the inductance due to the transformer primary or the tank-circuit inductance and capacitance. To demonstrate, when $\Delta i_C > 0$, the current rises, and the magnetic field set up around the coil stores some energy. The incremental change in current has the same sign as the current itself, the voltage drop across R_L is subtracted from E_2, and the collector voltage is brought down. In the circumstances, the emf of self-induction generated in the tank-circuit coil or the transformer winding opposes the current and impedes its rise. It also opposes the emf of the E_2 source, and the collector voltage falls below E_2.

When the current falls, the situation is reversed. The emf of self-induction changes sign and sustains the current. It is combined with the emf supplied by the E_2 source, and the collector voltage rises. In other words, the voltage drop across R_L changes sign and is added to, rather than subtracted from, E_2. This also follows from Eq. (6-29). When $\Delta i_C < 0$, the product $\Delta i_C R_L$ is added to E_2. When $\Delta i_C = -I_{C0}$, we obtain a maximum value

$$v_{CE\ \max} = E_2 + I_{C0}R_L$$

corresponding to point M.

Thus, the instantaneous collector voltage in a transformer-coupled or tank-circuit stage may be substantially higher than E_2. In other respects, the graphical constructions and calculations for the dynamic operation follow the same pattern, that is, as shown in Fig. 6-1 and using the equations given earlier.

Example. Let us determine the key quantities that characterize the operation of a transistor stage, using the numerical values given in Fig. 6-1. We choose the case when the signal source is operating as an equivalent constant-current generator (the Norton equivalent). The load line is constructed for $E_2 = 10$ V and $R_L = 2$ kΩ. Then

$$E_2/R_L = 10 \div 2 = 5 \text{ mA}$$

The operating region AB corresponds to

$$2I_{mB} = 80\ \mu\text{A}$$

$$2I_{mC} = 4.5 \text{ mA}$$

$$2V_{mCE} = 9 \text{ V}$$

Hence,

$$I_{mB} = 40\ \mu\text{A}$$

$$I_{mC} = 2.25 \text{ mA}$$

$$V_{mCE} = 4.5 \text{ V}$$

and

$$P_{out} = I_{mC}V_{mCE}/2 = (2.25 \times 4.5)/2 \approx 5 \text{ mW}$$

The Q-point defines

$$I_{B0} = 40\ \mu\text{A}$$

$$I_{C0} = 2.5 \text{ mA}$$

and

$$V_{CE0} = 5 \text{ V}$$

The power dissipated in the transistor is

$$P_{C0} = I_{C0}V_{CE0} = 2.5 \times 5 = 12.5 \text{ mW}$$

Using points A_1, B_1 and Q_1 on the input characteristics, we get

$$2V_{mBE} \approx 150 \text{ mV}$$

that is,

$$V_{mBE} = 75 \text{ mV}$$

and

$$V_{BE0} = 225 \text{ mV}$$

Now we are in a position to calculate the input power

$$P_{\text{in}} = 0.5 I_{m\text{B}} V_{m\text{BE}} = 0.5 \times 40 \times 10^{-6} \times 75 =$$
$$= 1.5 \times 10^{-3} \text{ mW}$$

and the input resistance

$$R_{\text{in}} = V_{m\text{BE}}/I_{m\text{B}} = 75 \times 10^3/40 = 1875\ \Omega$$

The respective gains are as follows:

$$k_I = I_{m\text{C}}/I_{m\text{B}} = 2.25 \times 10^3/40 = 56$$
$$k_V = V_{m\text{CE}}/V_{m\text{BE}} = 4.5 \times 10^3/75 = 60$$
$$k_P = k_I k_V = 56 \times 60 = 3360$$

or

$$k_P = P_{\text{out}}/P_{\text{in}} = 5 \times 10^3 \times 1.5 = 3330$$

The small discrepancy is a result of the inevitable inaccuracy of graphical calculations.

The power expended by the E_2 source is

$$P_0 = E_2 I_{\text{C0}} = 10 \times 2.5 = 25 \text{ mW}$$

and the efficiency is

$$\eta = P_{\text{out}}/P_0 = 5/25 = 0.2 = 20\%$$

Of course, in a low-power stage such as the one we are dealing with, the efficiency is of minor importance, and its calculation is only given as an example.

If the direct voltage is fed to the base from the E_2 source via a swamping resistor R_B, its resistance can be found by Ohm's law

$$R_\text{B} = (E_2 - V_{\text{BE0}})/I_{\text{B0}} =$$
$$= (10 - 0.225)/(40 \times 10^{-6})$$
$$\approx 0.25 \times 10^6\ \Omega = 250\ \text{k}\Omega$$

The procedures for constructing the load line and the associated calculations equally apply to CB circuits.

In all calculations of dynamic operation for a transistor, it is important to remember that the magnitude of the obtainable output power can be limited by a number of factors. Among other things, one may never exceed the safe limits for collector current, collector-to-emitter or collector-to-base voltage, and the power dissipated in the transistor chosen for the circuit in question. Figure 6-3 shows a shaded area – it represents the operating region for a CE circuit. From below this region is bounded by i_{CE0} (at $i_\text{B} = 0$). If it is important to minimize nonlinear distortion in the amplified signal, the operating region must further be bounded on the left (see the dashed line) – this will exclude the nonlinear portions of the characteristics. It should also be remembered that $P_{\text{C max}}$ must be brought down whenever there is a rise in ambient temperature and, as a consequence, in the temperature of the transistor case.

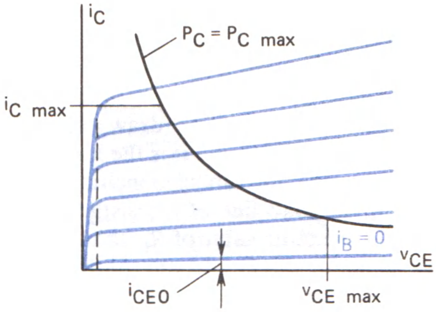


Fig. 6-3
Feasible region of a transistor

6-2 The Effect of Temperature on the Performance of Transistors

In operation, transistors are heated by the surroundings, extraneous heat sources (such as the nearby hot components), and the currents flowing through the transistors themselves. Variations in temperature have a marked effect on the performance of semiconductor devices. Among other things, an increase in temperature causes a rise in the conductivity of semiconductors and the currents flowing through them go up in magnitude. As has been shown in Sec. 3-3, the increase is especially noticeable in the reverse leakage current across a *p-n* junction. In transistors, this is the reverse collector leakage current. Any rise in this current entails changes in transistor characteristics. The situation can conveniently be traced by reference to the output characteristics shown for CB and CE circuits in Fig. 6-4.

The point can best be illustrated by a numerical example involving a germanium transistor for which $\beta = 100$ and $i_{\text{C0}} = 2\ \mu\text{A}$ at 20°C. Let the transistor be connected in a CB circuit and heated to 70°C, that is, by 50 degrees K above ambient. Since for germanium the reverse leakage current nearly doubles for every 10 degrees K of rise in temperature, in our case i_{C0} should increase by a factor of 2^5, that is, 32-fold. At $t = 70°\text{C}$ it will be 64 μA so that the net increase will be 62 μA. If we assume (rather arbitrarily)

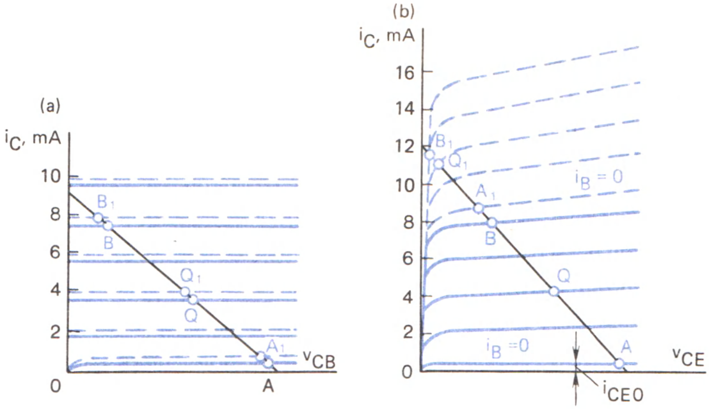


Fig. 6-4

Effect of temperature on the output characteristics of a transistor connected in (*a*) a CB circuit and (*b*) a CE circuit

that the alpha current gain is independent of temperature, then from the equality

$i_C = \alpha i_E + i_{CO}$

we may conclude that at i_E = const the collector current will go up likewise by 62 μA. Because i_C is a few milliamperes, this increase will only slightly affect the performance of the transistor.

The solid lines in the figure are the characteristics at $t = 20°C$, and the dashed lines, at $t = 70°C$. As is seen, in the CB circuit the characteristics are shifted slightly upwards. The operating point has somewhat shifted from Q to Q_1, and the new operating region, A_1B_1, only slightly differs from the region AB. In consequence, the amplification has remained nearly unchanged. Thus, the CB circuit is a temperature-stable configuration. Even with a temperature rise of tens of degrees the performance of the transistor in this circuit changes very little, and this is an important advantage.

The situation is entirely different when a transistor is connected in a CE circuit. In this configuration the reverse leakage current is i_{CEO}. It is approximately β times i_{CO}. In our example the figure at $t = 20°C$ is

$i_{CEO} \approx \beta i_{CO} = 100 \times 2 = 200\ \mu A$

At $t = 70°C$, this current is increased 32-fold and will be 6400 μA or 6.4 mA so that the increase will be 6.2 mA. It is seen from the equality

$i_C = \beta i_B + i_{CEO}$

that at i_B = const and β = const the collector current will build up in proportion to the rise in i_{CEO} (in our example, by 6.2 mA). Clearly, a heavy change like this cannot but drastically affect the output characteristics (Fig. 6-4*b*). The operating point Q and the operating region AB are now moved, respectively, to Q_1 and A_1B_1, and the dynamic operation in the amplification mode is completely upset. In our case, which is of course a hypothetical one, the portion A_1Q_1 of the operating region is greatly reduced and the portion B_1Q_1 is negligibly small. The gain is heavily decreased and the amplified signal will be strongly distorted because the positive half-cycle of input current remains practically unamplified. If we fail to apply the temperature compensation as has been explained in Sec. 4-5, the amplification provided by a CE circuit may be entirely unsatisfactory in the case of a temperature rise.

As is seen, the CE circuit has a low temperature stability – its performance is strongly affected by a rise in temperature and this is one of its major limitations in comparison with the CB configuration.

It is to be stressed that changes in temperature entail changes not only in the characteristics but also in all parameters of transistors. For example, assuming that the currents remain unchanged, a rise in temperature causes a certain increase in the h-parameters of the CE configuration. The parameters are affected more in the CE configuration but remain more stable in the CB connection. Temperature compensation serves to maintain the operating conditions of a transistor at a constant level, but it is unable to prevent changes in its parameters completely.

6-3 The Frequency Behaviour of Transistors

A rise in frequency leads to a fall in the amplification supplied by a transistor. Two factors may be held responsible for this occurrence. Firstly, the collector junction capacitance C_C produces a detrimental effect at high frequencies. A simple way to get insight into the matter is by reference to the Norton (current-generator) equivalent circuit shown for the CB connection in Fig. 6-5. At low frequencies, C_C presents a very high reactance; also r_C is very large (as a rule, r_C is many times R_L), so we may take it that all of the current αI_{mE} is flowing to the load resistor or that $k_I \approx \alpha$. At some high frequency, however, the C_C reactance drops to a relatively low value and a sizeable proportion of the current supplied by the generator divides into this reactance and a smaller current flows through R_L. In consequence, k_I, k_V, k_P, and output power suffer a reduction.

If we assume that the frequency is tending to infinity, the capacitive reactance $1/\omega C_C$ will tend to zero – C_C will present an effective short-circuit to the generator, all of its current αI_{mE} will flow through C_C, and no current at all will be flowing in the load resistor. We would get the same result if we use a Thévenin (voltage generator) equivalent circuit.

The emitter junction capacitance C_E likewise falls off with a rise in frequency, but it is always shunted by the low emitter-junction resistance r_E, and so its detrimental effect may be felt only at very high frequencies when $1/\omega C_E$ is comparable in magnitude with r_E.

Basically, the effect of C_E consists in that its reactance decreases with increasing frequency with the result that its shunting action on r_E becomes stronger, and the alternating voltage which controls the collector current is brought down. Naturally, the gain is reduced in proportion. If the frequency is tending to infinity, the reactance $1/\omega C_E$ will tend to zero, and the voltage across the emitter junction will likewise tend to zero. In practice, C_C shunted by the very large collector-junction resistance r_C produces so strong an effect even at less higher frequencies that the use of a transistor at higher frequencies where C_E could be noticeable becomes unwarranted. Therefore, the effect of C_E need not be considered in many cases.

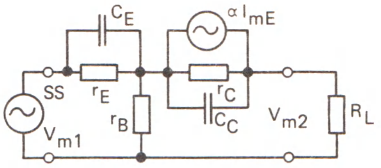


Fig. 6-5
Equivalent circuit of a transistor, with its junction capacitances included

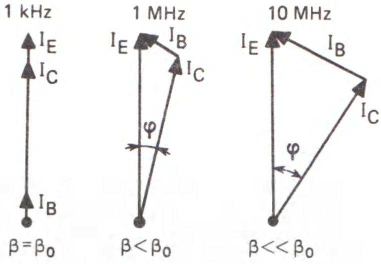


Fig. 6-6
Phasor diagrams for transistor currents at different frequencies

To sum up, the effect of C_C in operation at high frequencies consists in that it brings about a reduction in the alpha and beta current gains, α and β.

The other cause of the reduction in gain at higher frequencies lies in that the alternating collector current lags behind the alternating emitter current in phase. This lag is due to the delay caused by the movement of carriers from the emitter to the collector through the base and also by the carrier storage in the base. Carriers (say, electrons if we take an *n-p-n* transistor) move in the base by diffusion, and so their velocity is not very high. Their transit time through the base, τ_t, in ordinary transistors is 10^{-7} s, that is, 0.1 μs or even less. Of course, this is a very short time interval, but at frequencies of several units or tens of megahertz and higher it is comparable with the period of oscillation and causes a noticeable phase shift between collector and emitter currents. This phase shift at high frequencies is responsible for a rise in the alternating base current which in turn brings down the beta current gain.

It is most convenient to trace this occurrence by reference to the phasor diagrams shown in Fig. 6-6. The first corresponds to a low frequency, say 1 kHz, at which all currents are practically in phase because τ_t is only a tiny fraction of the oscillation period. At low frequencies, the beta current gain takes on its maximum value β_0. At higher frequencies, say 1 MHz, the fact that I_C lags behind I_E in time by τ_t gives rise to a marked phase shift φ between the two currents. Now I_B is a geometrical (that is, phasor), and not an algebraic, difference between I_E and I_C and, in consequence, it builds up appreciably. Therefore, even if I_C remained unaffected by C_C, the beta current gain would be smaller than β_0 all the same. At a still higher frequency, say 10 MHz, the phase shift becomes greater, I_B takes a higher value, and the beta current gain falls.

Thus, a rise in frequency causes β to decrease substantially more than α. The alpha current gain falls solely due to the effect of C_C, and the beta current gain is additionally affected by the phase shift between I_C and I_E owing to the transit time of carriers across the base. It follows then that the CE circuit has a poorer frequency performance in comparison with the CB connection.

It is customary to set the allowable limit of fall in the two current gains by 30% from their values α_0 and β_0 at low frequencies (usually, a frequency of 1 kHz is taken). The frequencies at which the gain falls so that $\alpha = 0.7\alpha_0$ and $\beta = 0.7\beta_0$, are called the *common-base current-gain cutoff frequency* and the *common-emitter current-gain cutoff frequency*, respectively for CB and CE circuits. In this text, they will be labelled as f_α and f_β. Outside the Soviet Union, they are often designated as f_{ab} and f_{ae}. Their more rigorous definition may be stated as follows:

(1) f_{ab} is the common-base current-gain cutoff frequency or the frequency at which h_{fb} (that is, the common-base a. c. forward current gain or alpha) has decreased to a value 3 dB below h_{fbo} or α_0, that is, $h_{fb} = 0.707 \times h_{fbo}$.

(2) f_{ae} is the common-emitter current-gain cutoff frequency or the frequency at which h_{fe} (that is, the common-emitter a. c. forward current gain or beta) has decreased to a value 3 dB below h_{feo} or β_0, that is, $h_{fe} = 0.707 \times h_{feo}$.

Because the beta current gain decreases at a far higher rate than the alpha current gain, f_β is markedly lower than f_α. We may write

$$f_\beta \approx f_\alpha/\beta \tag{6-30}$$

Figure 6-7 shows an approximate plot illustrating how the alpha and beta current gains of some transistor fall off with rising frequency laid off on a log scale. For convenience, the ratios α/α_0 and β/β_0 are laid off on the vertical axis instead of α and β.

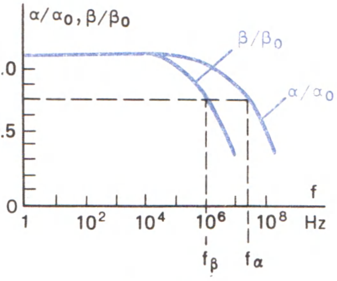


Fig. 6-7

Reduction in the alpha and beta gains with rising frequency

In addition to the two current-gain cutoff frequencies f_α and f_β, it is usual to quote for each transistor its *maximum frequency of oscillation*, f_{max}, which is defined as the frequency at which the power gain k_P drops to unity (or 0 dB). Obviously, at $f < f_{max}$ where $k_P > 1$, a given transistor may be used in a self-excited oscillator. If, on the other hand, $k_P < 1$, no oscillations will be generated and sustained.

Sometimes, the design equations include also what is known as the *gain-bandwidth product*. It is designated as f_T and defined as the frequency at which the common-emitter forward current gain k_I is equal to unity, that is, when a CE circuit ceases to amplify current.

Changes in the alpha and beta current gains are not the only events that happen in operation at high frequencies. The effect of junction capacitances, carrier storage, and transit-time effects also change the primary (T or r) parameters of a transistor at high frequencies so much that they no longer act as pure resistances. The other parameters are affected as well.

The frequency response of transistors can be improved, that is, their cutoff frequencies f_α and f_β can be boosted by bringing down the collector junction capacitance C_C and the transit time τ_t. Unfortunately, any reduction in C_C by decreasing the surface area of the collector junction inevitably leads to a lower absolute maximum current rating, that is, to a lower absolute maximum power rating.

The value of C_C can somewhat be reduced by decreasing the impurity concentration in the collector. As a result, the collector junction widens, and this amounts to an increase in the spacing between the plates of an equivalent capacitor. As a result, the capacitance is reduced and, also, since the junction is now wider, a higher voltage would be needed to cause its breakdown, and so the absolute maximum power rating of such a transistor may be set at a higher value. Unfortunately, this entails an increase in the resistance presented by the collector region and, in consequence, in the power dissipation there, which is an obvious disadvantage, especially for high-power transistors. The transit time can be reduced by making the base very thin and by imparting a greater velocity to the carriers that move through the base. With a thinner base, however, one has to lower V_{CB}, or else a punch-through might occur as the collector junction widens. Under diffusion, electrons are far more mobile than holes. Therefore, with all other conditions being equal, *n-p-n* transistors are able to operate at far higher frequencies than *p-n-p* transistors. Higher current-gain cutoff frequencies can also be obtained by using semiconductors in which the carriers have a greater mobility. A shorter transit time is also obtained in transistors with an electric field set up in the base region so as to accelerate the carriers. In more detail, the design and operation of high-frequency transistors are covered in Sec. 6-7.

6-4 Transistors as Switches

Transistors are widely used in pulse circuitry owing to their special behaviour in what may be called the *switching mode of operation*.

We will examine a *switching transistor* by reference to its output characteristics for the CE circuit configuration. Let the collector lead contain a load resistor R_L. The applicable load line appears in Fig. 6-8. Before an input current or voltage pulse arrives at the input of this circuit, the transistor is turned off (it is said to be in the OFF state or *at cutoff*). This condition corresponds to point Q_1. A small current is now flowing in the collector circuit (this is the transfer leakage current i_{CE0}), and so this circuit may be assumed to be open. Nearly all of the supply voltage E_2 is impressed on the transistor.

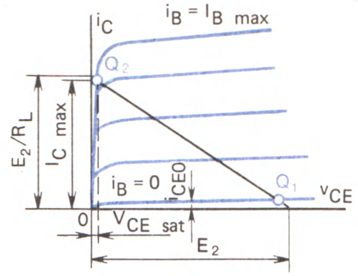


Fig. 6-8

Finding transistor parameters in pulsed work from output characteristics

When we apply a current pulse $I_{B\,max}$ to the input, the transistor is moved into the *saturation region* and is operating at point Q_2. As a result, there appears a collector current pulse $I_{C\,max}$ very close in magnitude to E_2/R_L. Sometimes, it is called the *saturation current*. At saturation, a transistor is in effect a closed switch and nearly all of E_2 is dropped across R_L, with a very small remainder, called the *saturation voltage* $V_{CE\,sat}$, left across the transistor; its value is a few tenths of a volt.

Although v_{CE} does not change sign at Q_2, it becomes a forward voltage across the collector junction itself, and so Q_2 represents the saturation region. This can best be illustrated by the following example. Suppose we have an *n-p-n* transistor in which $V_{CE\,sat} = 0.2$ V and $V_{BE} = 0.6$ V. Then the collector-to-base voltage will be $V_{CB} = 0.2 - 0.6 = -0.4$ V which indicates that a forward voltage of 0.4 V exists across the collector junction.

Of course, if the input current pulse is smaller than $I_{B\,max}$, the collector current pulse will likewise be smaller. On the contrary, an increase in the base current pulse over and above $I_{B\,max}$ will not practically produce an increase in the output current pulse. Thus, the maximum attainable value of collector current pulse is

$$I_{C\,max} \approx E_2/R_L \qquad (6\text{-}31)$$

In addition to $I_{C\,max}$, $I_{B\,max}$ and $V_{CE\,sat}$, the switching mode of operation is characterized by the current gain B which, in contrast to the beta current gain, is defined in terms of the ratio between the currents at point Q_2

$$B \approx I_{C\,max}/I_{B\,max} \qquad (6\text{-}32)$$

rather than in terms of incremental changes in the respective currents (the respective notation used outside the Soviet Union is h_{FE} and h_{fe}). In other words, the beta current gain is a small-signal parameter while B is a direct-current (or, rather, large-signal) parameter, so there is a

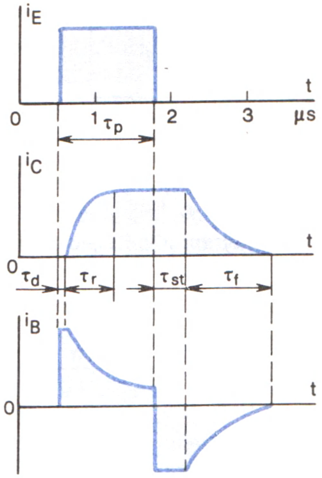


Fig. 6-9

Corruption of the current pulse waveform by a transistor

difference in magnitude between the two.

Still another parameter associated with switching transistors is the *saturation resistance* defined as

$$R_{\text{sat}} = V_{\text{CE sat}}/I_{\text{C max}} \tag{6-33}$$

For switching transistors, R_{sat} is usually units or, sometimes, tens of ohms.

The operation of a switching transistor in a CB circuit is similar to that in the CE configuration we have just discussed.

If the input pulse duration τ_p is many times the storage time τ_{st} of a transistor, the output current pulse will have about the same duration and shape as the input pulse. In the case of short pulses, that is, when τ_p is several microseconds or even less, the waveform of the output current pulse may be heavily distorted and its duration increased.

As an example, Fig. 6-9 shows plots of a short input current pulse which has a rectangular waveform and an output current pulse for a transistor with the CB connection. As is seen, the collector current pulse has a delay τ_d (*delay time*) owing to the finite transit time of carriers through the base. This current builds up gradually for a time τ_r (the *pulse rise time*) which accounts for an appreciable part of τ_p. This gradual rise is associated with the carrier storage in the base. Also, the carriers injected into the base at the start of the input current pulse differ in travel velocities and they do not reach the collector all at the same time. The time $\tau_d + \tau_r$ is the *turn-on time*, τ_{on}, of a transistor. After the input pulse has ceased, the collector current i_C keeps flowing for some time equal to the storage time τ_{st}, and then gradually decays during a time interval called the *pulse fall* (or *decay*) *time*, τ_f. The time interval $\tau_{st} + \tau_f$ constitutes the *turn-off time* of a transistor, τ_{off}. The net result is that the collector current pulse markedly differs from a rectangular pulse in shape and is stretched in time as compared with the input pulse. In consequence, a longer time is required for the collector circuit to be turned on and off and for the transistor to remain in the ON state. In other words, carrier storage in the base impedes a fast turn-on and a fast turn-off for a transistor or, which is the same, a switching transistor has a limit to its speed of operation.

Figure 6-9 also shows a plot of the base current, constructed on the basis of the relation

$$i_B = i_E - i_C$$

As is seen, this current does not have a simple waveform.

Timing current diagrams for the CE configuration may be plotted similarly to those shown in Fig. 6-9 for the CB connection.

Switching transistors intended to handle short pulses must have small junction capacitances and a narrow base. As a rule, they are low-power drift transistors (see Sec. 6-7). As a way of cutting down the storage time, it is usual to add to the base region a small amount of an impurity which speeds up the recombination of stored carriers (this impurity may be, say, gold).

6-5 Frequency Changing by Semiconductor Devices

In a wider sense, frequency changing applies to any change in frequency. For example, rectification converts an alternating current at frequency *f* to a direct current at zero frequency. In oscillators, a direct current of zero frequency is converted to an alternating current of a desired frequency. In this section we will deal with frequency changing which has as its objective to transform a signal at one frequency other than zero to a signal at another frequency likewise other than zero.

The frequency changer or converter shown in the block diagram of Fig. 6-10 consists of a mixer, a source which supplies an a. c. voltage v_s at frequency f_s which is to be converted, and a second source, called the *local oscillator*, which supplies an auxiliary a. c. voltage v_{lo} at frequency f_{lo}. As often as not, the mixer and the low-power local oscillator are combined in a single tube. The signal appearing at the output of the mixer (or the frequency converter, as the case may be) is at what is termed the *intermediate frequency*, f_i, usually abbreviated as IF.

A mixer must be a nonlinear device. If it were a linear device, the two waveforms fed to it would simply be added together. For example, if these two waveforms are at closely spaced, but not multiple frequencies (Fig. 6-11*a* and *b*), the result will be beats – a complex waveform whose frequency is oscillating within certain limits about its average (or arithmetic mean) value

$$f_{av} = (f_s + f_{lo})/2 \qquad (6\text{-}34)$$

and its amplitude is varying at a frequency equal to a difference frequency, $f_{lo} - f_s$ (Fig. 6-11*c*).

These beats do not contain a component at a new frequency. If, on the other hand, we detect (rectify) the beats, the nonlinearity of this process will extract a component at the intermediate frequency, $f_i = f_{lo} - f_s$ (Fig. 6-11*d*).

The signal appearing at the output of a mixer or frequency converter is a complex waveform containing components at different frequencies. A general equation for these frequencies is

$$f = |mf_s \pm nf_{lo}| \qquad (6\text{-}35)$$

where m and n are any integers, including zero.

When $m = n = 0$, the result is $f = 0$. This means that a direct component exists at the frequency converter output. On setting $m = 1$ and $n = 0$ or $m = 0$ and $n = 1$, we find that the output signal includes components at frequencies f_s and f_{lo}. When $m = n = 1$, the output waveform contains new components, one at the difference frequency $f_s - f_{lo}$ and the other at the sum frequency $f_s + f_{lo}$. One of them is ordinarily utilized as the converted frequency. Other values of m and n will produce further new frequencies.

All of these new frequencies are combinations of f_s and f_{lo} and of their harmonics. Quite aptly, they are sometimes called *combination frequencies*. More often, they are referred to as *intermodulation products* because they arise from the process of intermodulation in which one frequency 'modulates' the other. By choosing a suitable auxiliary, or local-oscillator, frequency f_{lo}, we can derive any desired new frequency.

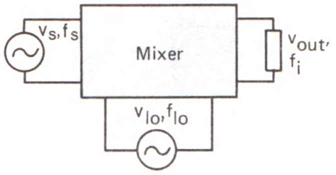


Fig. 6-10

Generalized diagram of a frequency changer

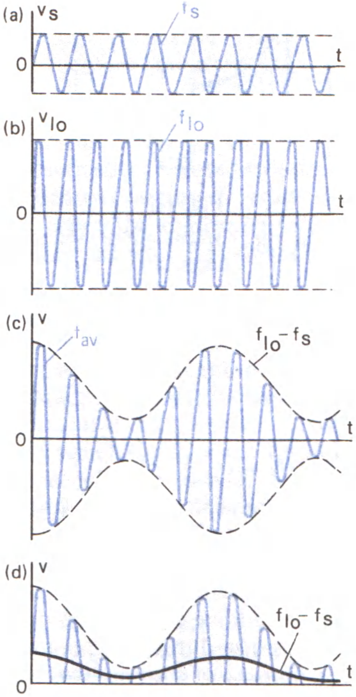


Fig. 6-11

Explaining frequency conversion

The new frequencies produced by conversion include the harmonics of the original waveforms at frequencies $2f_s$, $3f_s$, ... and $2f_{lo}$, $3f_{lo}$, ... Provision of two waveforms is not mandatory for harmonics to arise – they can result from a single waveform subjected to nonlinear distortion.

As a rule, intermodulation products decrease progressively in amplitude as their numbers m and n increase. Therefore, the new or intermediate frequency is most often the difference frequency, although the sum frequency is also used sometimes. Higher intermodulation products are used but seldom.

In radio receivers, frequency conversion is mostly carried out in such a way that one and the same IF is produced even though the sending stations may operate at different frequencies. The use of a single IF makes it possible to assure a high gain and a high selectivity which remain almost constant over the entire range of received frequencies. Also, since they are to amplify a single frequency, the amplifying stages can readily be designed for a more stable operation and to have a simpler circuitry than stages intended to handle several frequencies or a range of frequencies.

Radio receivers and instruments most often use the difference frequency as the IF, with the local-oscillator frequency f_{lo} usually chosen to be higher than the signal frequency f_s. This choice of the local-oscillator frequency is mandatory if the intermediate frequency is to be higher than the signal frequency. Also, if the conversion is carried out over a range of frequencies, a smaller relative change in f_{lo} will be required when $f_{lo} > f_s$, and a simpler circuitry may be used for the local oscillator. For example, if the signal frequency varies from 500 to 1000 kHz and the IF must be 250 kHz, then with the local-oscillator frequency chosen to lie below the signal frequency, it would have to be varied from $500 - 250 = 250$ kHz to $1000 - 250 =$ $= 750$ kHz, that is, by a factor of three. When, on the other hand, $f_{lo} > f_s$, the local-oscillator frequency needs to be varied from $500 + 250 =$ $= 750$ kHz to $1000 + 250 = 1250$ kHz, or by a factor of less than two.

In our subsequent discussion, the intermediate frequency f_i will be taken to mean the difference frequency.

The name 'intermediate frequency' has come about historically. The point is that in supersonic heterodyne receivers, most commonly called superheterodyne receivers or simply superhets, all incoming signals are translated so that the carrier frequency is changed to a lower value constant for all carriers and most of the predetector amplification is then effected at this frequency as an *intermediate* step in reception. The IF is then detected to yield a signal at an audio frequency F. (In some of the existing superhet receivers, the IF lies above f_s.) Therefore the term 'intermediate frequency' ought to be construed in the sense that the IF amplifier stages are placed in a superhet receiver between the RF stages handling the incoming frequencies f_s and the AF stages delivering the audio frequencies F.

Various nonlinear devices may be used for frequency conversion. For example, the frequency converters (mixers) of UHF and SHF receivers use semiconductor diodes. Transistors are used for frequency conversion in the UHF, VHF and lower frequency bands.

Frequency conversion occurs as follows. The mixer, which is a nonlinear device, is fed voltages at f_s and f_{lo}. The current flowing through the device is thus made to pulsate at these two frequencies, and the nonlinearity of the device gives rise to components at combination frequencies. The output resonant circuit is tuned to resonate at one of them, usually at the difference frequency, $f_i = f_{lo} - f_s$. It presents a high impedance only to the current at that resonance frequency, so it delivers an amplified voltage only at f_i.

In this way, the output resonant circuit extracts the intermediate frequency.

In frequency converters it is vitally important to avoid coupling between the signal circuits and the local-oscillator circuits. As a rule, both contain resonant circuits. If no measures were taken to the contrary, coupling between them might allow the resonant circuits to affect one another, thereby disturbing the tuning. Also, the stability of the local oscillator would be impaired and, if there is no r.f. amplifier, stray radiations originating in the local oscillator might be emitted by the antenna (known as the *heterodyne whistle*). Furthermore, a strong incoming signal, such as one from a nearby radio station, might break through to the local oscillator and cause it to generate the same frequency despite the fact that the resonant circuit of the local oscillator has been tuned to resonate at $f_{lo} = f_s + f_i$. The local oscillator is then said to be in a *lock-in* with the incoming signal. This is an undesirable occurrence because when the signal and local-oscillator frequencies are the same, no intermediate frequency can be produced, and no reception is possible.

Two simple frequency converters are shown in the circuit diagrams of Fig. 6-12. In the circuit of Fig. 6-12*a*, the mixer diode accepts a voltage at the signal frequency f_s and a voltage at the local-oscillator frequency f_{lo}. The $L\,C$-circuit is tuned to resonate at a new (intermediate)

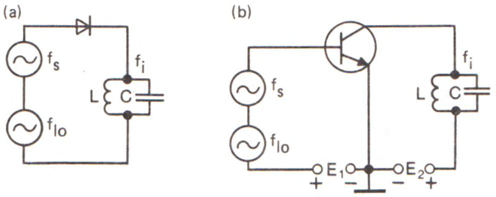


Fig. 6-12
Frequency conversion circuits

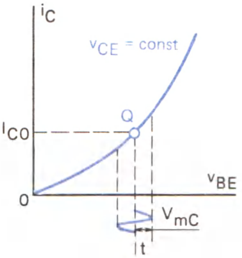


Fig. 6-13
Transfer characteristic of a transistor

frequency. In the CE circuit of Fig. 6-12*b*, the two voltages are fed to the base of a bipolar transistor. A similar circuit can be set up, using a FET connected in the common-source (CS) configuration*. A drawback common to both arrangements is that no measures are taken to avoid coupling between the signal source and the local oscillator. This coupling may somewhat be reduced in the bipolar-transistor circuit by placing the local oscillator in the emitter lead (or in the source lead in the case of a FET). Quite a number of arrangements have been developed which effectively decouple the signal and local-oscillator circuits.

Let us go through the frequency conversion process in a bipolar-transistor circuit with the aid of simple mathematical relations. The control (or dynamic transfer) characteristic, $i_C = f(v_{BE})$, of the transistor is shown in Fig. 6-13. The signal voltage $v_s = V_{ms} \sin \omega_s t$ and the local-oscillator voltage $v_{lo} = V_{mlo} \sin \omega_{lo} t$ are applied to the transistor base.

The collector current may be written as

$$i_C = I_{CO} + g_m V_{mlo} \sin \omega_{lo} t + g_m V_{ms} \sin \omega_s t \quad (6\text{-}36)$$

where $g_m = y_{21e}$ is the forward transconductance of the transistor, defined as $\Delta i_C / \Delta v_{BE}$.*

Because a transistor is a nonlinear device, its dynamic transfer characteristic is likewise nonlinear, and its forward transconductance itself is a function of voltage. Let us deem that the forward transconductance varies under the influence of the local-oscillator voltage. For simplicity we assume that the characteristic is quadratic (a parabola); if so, the forward transconductance will vary linearly as

$$g_m = g_{m0} + g_{m\,m} \sin \omega_{lo} t \quad (6\text{-}37)$$

where g_{m0} = initial forward transconductance at the *Q*-point (in the quiescent state)

$g_{m\,m}$ = peak value of the forward transconductance.

On substituting the two transconductances in Eq. (6-36) and carrying out some manipulations, we obtain

$$\begin{aligned} i_C &= I_{CO} + (g_{m0} + g_{m\,m} \sin \omega_{lo} t)\, V_{mlo} \sin \omega_{lo} t \\ &\quad + (g_{m0} + g_{m\,m} \sin \omega_{lo} t)\, V_{ms} \sin \omega_s t \\ &= I_{CO} + g_{m0} V_{mlo} \sin \omega_{lo} t + g_{m\,m} V_{mlo} \sin^2 \omega_{lo} t \\ &\quad + g_{m0} V_{ms} \sin \omega_s t + g_{m\,m} V_{ms} \sin \omega_{lo} t \sin \omega_s t \end{aligned} \quad (6\text{-}38)$$

As trigonometry tells us,

$$\sin^2 \alpha = (1 - \cos 2\alpha)/2$$

and

$$\sin \alpha \sin \beta = \cos(\alpha - \beta)/2 - \cos(\alpha + \beta)/2$$

Therefore, we may write

$$\begin{aligned} i_C &= I_{CO} + g_{m0} V_{mlo} \sin \omega_{lo} t + g_{m\,m} V_{mlo}/2 \\ &\quad - g_{m\,m} V_{mlo} \cos 2\omega_{lo} t/2 + g_{m0} V_{ms} \sin \omega_s t \\ &\quad + g_{m\,m} V_{ms} \cos(\omega_{lo} - \omega_s)/2 \\ &\quad - g_{m\,m} V_{ms} \cos(\omega_{lo} + \omega_s)\, t/2 \end{aligned} \quad (6\text{-}39)$$

Thus, the nonlinear properties of a transistor are responsible for the fact that the collector current contains a direct component, the local-oscillator frequency, the signal frequency, the second local-oscillator harmonic, and new components at the difference and sum frequencies. It is one of these new components that may be used as the intermediate frequency. Because the characteristic of a transistor is usually more complex

* FETs are discussed in Sec. 7-1.

* There is no consensus on the symbol for this quantity. Most authors give it the symbol g_m by analogy with electron tubes. Others symbolize it as Y_{fe}, still others as g_{fe} or even as g_{21}.– *Translator's note.*

than a parabola, the forward transconductance is a nonlinear function of voltage. This gives rise to further components at various combinations of frequencies, that is, to ever more intermodulation products.

In Eq. (6-39), the currents at the difference and sum frequencies have the same amplitude equal to

$$I_{mC\,IF} = g_{m\,m} V_{ms}/2 \tag{6-40}$$

The ratio $I_{mC\,IF}/V_{ms} = g_c$ is given the special name of *conversion transconductance*; it shows how effective a given stage is in terms of frequency conversion. It follows from Eq. (6-40) that

$$g_c = g_m/2$$

That is, g_c does not exceed half the maximum change in the forward transconductance at the Q-point. As V_{mlo} is raised, so is the change in the transconductance, leading to a higher value of g_m and, in consequence, of g_c.

The IF voltage appearing at the stage output is given by

$$V_{m\,IF} = I_{mC\,IF} R_L \tag{6-41}$$

where R_L is the load resistance, that is, the opposition presented by the output resonant circuit tuned to resonate at the intermediate frequency.

On re-writing, we have

$$V_{m\,IF} = g_{m\,m} R_L V_{ms}/2 = g_c R_L V_{ms} \tag{6-42}$$

Hence, the gain of a transistor amplifying stage is

$$k = V_{m\,IF}/V_{ms} = g_c R_L \tag{6-43}$$

6-6 Inherent Noise in Transistors

At high amplification, a pair of headphones or a speaker plugged into the output of an amplifier or receiver will give out a characteristic hissing sound even when no valid signal is fed to the input. The same kind of noise can be heard from any radio receiver on shorting together its input terminals so as to avoid picking up any extraneous signals. The higher the amplification, the stronger this inherent receiver noise.

As studies have shown, the currents and voltages in any electric circuit are always fluctuating at random owing to the thermal agitation of electrons. As the temperature around a circuit rises, the fluctuations build up.

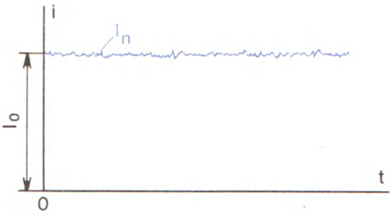


Fig. 6-14
Current fluctuations

All currents (base, collector and emitter) of a transistor are subject to fluctuations. After amplification, they produce, when received by ear, *fluctuation* or *random noise* which is inherent (or intrinsic) to the device. Therefore, we may call it *inherent noise* even when incoming signals are not converted to audible signals.

Any direct current is not exactly constant in magnitude, but contains what may be called an alternating noise component I_n in addition to its direct component I_0. The point is that owing to thermal motion the number of electrons passing through the cross-sectional area of any conductor during small equal time intervals is varying continuously rather than remains constant. Current fluctuations are illustrated in Fig. 6-14 where the noise current is shown on an exaggerated scale (ordinarily, it is vanishingly small in comparison with I_0).

It has been postulated theoretically and verified by experiment that noise current is a sum of alternating sinewave components at frequencies ranging from zero to extremely high. Any amplifier (or any other device) can, however, pass frequencies within some particular range. Therefore, the amplifier output delivers only the noise components whose frequencies fall within this frequency range called the *bandwidth* of the amplifier, B. The larger the bandwidth, the greater is the proportion of noise that will be passed by the device or circuit.

The inherent noise of transistors limits the sensitivity of radio receivers or other devices intended to detect, amplify and/or measure weak signals. If valid signals are weaker than inherent noise, they will be drowned or masked by noise in part or even completely, and poor or no reception will result.

In any resistor, the fluctuations that take place in it give rise to a noise emf. The rms value, E_n, of noise emf generated in a resistor or any circuit of resistance R is given by *Nyquist's equation*

$$E_n = (4kTRB)^{1/2} \tag{6-44}$$

where E_n = rms value of noise emf
k = Boltzmann's constant, 1.38×10^{-23} J K^{-1}
T = absolute temperature of the resistor

For practical calculations and at room temperature, this equation takes the form

$$E_n \approx (RB)^{1/2}/8 \tag{6-45}$$

where E_n is in microvolts, R in kilohms, and B in kilohertz.

For example, with $R = 40$ kΩ and $B = 10$ kHz, we have

$$E_n = (40 \times 10)^{1/2}/8 = 20 \div 8 = 2.5\ \mu\text{V}$$

It is to be noted that inherent noise is produced in semiconductor diodes as well, but it has to be taken into account only when the diodes are used in the first stage of microwave receivers.

There are several sources of noise in a transistor. Some of them are described in brief below.

Thermal noise is due to the thermal agitation of electrons in any resistor. Because any of the transistor regions possesses a resistance, there will occur a noise voltage. Since the emitter and collector regions have a relatively low resistance, thermal noise in a transistor mainly comes from the base resistance r_B, the more so that it is connected in the input circuit and its noise is amplified by the transistor itself.

Shot noise (also known as *Schottky noise*) comes about from fluctuations in the injection and extraction of carriers into and from the emitter and collector junctions.

Partition noise occurs due to random fluctuations in the division of the emitter current between the base and collector.

Recombination noise has its origin in the random fluctuations of carrier (electron and hole) recombination.

There are also further causes of noise due to the random fluctuations in the leakage currents in the surface layers of semiconductors and some other factors. The result is what is known as *excess* or *flicker noise*.

Noise power or voltage increases with increasing bandwidth within which the effect of noise manifests itself.

The noise characteristics of transistors are often specified in terms of the so-called *noise figure*, F_n. As with two-ports having a generator connected to the input terminals, it is defined as follows.

The influence of noise is always characterized by the ratio of signal power P_s to noise power P_n (the signal-to-noise-power ratio). At the output, this signal-to-noise-power ratio is always smaller than it is at the input because at the output the two terms appear amplified by a factor of k_P, but the transistor adds its inherent noise power $P_{n\,tr}$ to the noise power. The noise figure shows how many times the available signal-to-noise-power ratio at the input (signal-generator) terminals is greater than the available signal-to-noise-power ratio at the output

$$F_n = \frac{P_{s\,in}/P_{n\,in}}{P_{s\,out}/P_{n\,out}} \tag{6-46}$$

Or, in decibels,

$$F = 10 \log_{10} F_n \tag{6-47}$$

An F_n of 10, 100 and 1000 is respectively equal to an F of 10, 20, and 30 dB.

State-of-the-art transistors have an F of about 3 to 30 dB (on average, 10-20 dB). The noise figure is usually specified in data sheets for a frequency of 1 kHz and a temperature of 20°C.

The magnitude of transistor noise depends on the transistor parameters, mode of operation, and the internal resistance of the signal source, R_{ss}. The lower the alpha current gain of a transistor, the higher its noise. The point is that a decrease in the alpha current gain is accompanied by a rise in the base current, so it produces across r_B a higher noise voltage which is further amplified by the transistor. Also, as the alpha current gain falls, more carriers recombine in the base, and the recombination is one of the causes of inherent noise in transistors.

Noise also goes up with increasing r_B and i_{CO}. The kind of semiconductor material also affects the noise level. For example, silicon transistors are more noisy than germanium devices. A reduction in v_{CB} and i_E can reduce noise, but to a certain definite limit only because very small values of v_{CB} and i_E result in a low value of the alpha current gain, and this might lead to a build-up in noise. For noise to be kept to a minimum, R_{ss} must be held at some optimal value, usually several hundred ohms. A rise in temperature brings about a sudden increase in the inherent noise of transistors. Theory and

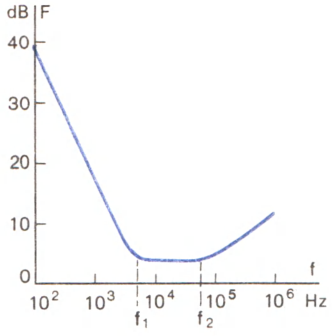


Fig. 6-15

Noise figure of a transistor as a function of frequency

experience show that, with all other conditions being equal, transistor noise has about the same magnitude for all the three basic circuit configurations: CE, CB, and CC.

Noise is distributed with frequency other than uniformly. It is seen from Fig. 6-15 that at medium frequencies F has a minimal and approximately constant value. The lower frequency limit f_1 of this range is several kilohertz. At frequencies below f_1 there is an increase in excess (or flicker) noise, and as a result F is increased. At frequencies above f_2 there is an increase in F owing to a reduction in the alpha current gain. The value of f_2 may be several hundred kilohertz or even more. As f_α goes up, so does f_2 which is by a factor of about $\beta^{1/2}$ lower than f_α.

All of the above relations are taken into account in the manufacture of special-purpose low-noise transistors for the early stages of amplifiers and radio receivers. For noise to be a minimum, such transistors operate at reduced values of v_{CB} and i_E, and their temperature must be kept low. These transistors have high values of α and f_α, but low values of r_B and i_{CO}.

Good transistors are less noisy than tubes at medium frequencies, but are more noisy at low and high frequencies.

6-7 Basic Types of Bipolar Transistors

Existing transistors may be classed by technology, material, service, frequency range, power output, and some other factors. As already noted, point-contact transistors, although they were invented first, are no longer used. The semiconductor materials used in the manufacture of transistors are only germanium and silicon, but it appears that some other materials will also be used in the future. In terms of power dissipation, transistors are classed into low-power, medium-power, and high-power, the respective figures of $P_{C\,max}$ being up to 0.3 W, from 0.3 to 1.5 W, and over 1.5 W. In terms of frequency limit (gain-bandwidth product), transistors are classed into low-frequency types (capable of operation at up to 3 MHz), medium-frequency (from 3 to 30 MHz), and high-frequency (over 30 MHz).

In most transistors, the prevailing mechanism is the injection of carriers in the emitter junction, but there is a group of transistors which operate without injection. Among them are field-effect transistors, or FETs (see Chap. 7).

The most commonly used devices are bipolar transistors with two *p-n* junctions. Their operation has been examined in detail in the previous sections. They may be classed into *drift transistors* in which the transfer of minority carriers through the base is mainly by drift, that is, under the influence of an accelerating electric field, and *diffusion transistors* in which this transfer is mainly by diffusion. Diffusion transistors ought not to be confused with *diffused transistors*. The term 'diffusion transistor' defines a transistor relying on the dissemination of carriers for the flow of current. A *diffused transistor* is that in which the emitter and collector junctions are both formed by the diffusion technology.

It is to be noted that, given a heavy injection of carriers from the emitter, an electric field will be set up in the base even in diffusion transistors, so the transfer of carriers there will not be by diffusion alone. By the same token, some carriers are transferred by diffusion even in the base of drift transistors despite the fact that the main mechanism of transfer is drift.

In diffusion transistors, the impurity concentration is the same throughout the base region. As a result, no electric field is set up, and the carriers diffuse from emitter to collector. The velocity of motion due to diffusion is lower than that due to drift in an accelerating field. In consequence, diffusion transistors can operate at lower frequencies than drift devices.

In drift transistors, the electric field of the base accelerates minority carriers as they move towards the collector. In consequence, they have a higher frequency limit and a greater current gain. The electric field is produced in the base due to the fact that the impurity concentration is not the same at various points in the base. This can be produced when *p-n* junctions are fabricated by the diffusion process. As already noted, such devices are called diffused transistors.

The electric field in the base of a diffused transistor is brought about as follows. Let, as an example, the base contain a donor impurity so that *n*-type (or electron) conduction is produced. If the concentration of this impurity near the emitter junction is greater than it is near the collector junction, there will be a proportionate difference in the concentration of majority carriers (electrons in our case) in the base. It will be higher at the emitter junction. Owing to this concentration gradient, some electrons move to the region of a lower concentration, that is, towards the collector junction (Fig. 6-16). The base comes by a difference of potential (with the " – " side near the collector and the " + " side near the emitter) which sets up an electric field. This field retards the majority carriers, that is, impedes any further travel of electrons. At equilibrium, the effect of the potential difference on electrons is counteracted by the difference in concentration (the concentration gradient), and the base acquires an electric field which accelerates the minority carriers (holes) injected from the emitter.

Diffusion transistors may have alloyed or fused junctions formed by the same technology as diodes. These are *alloyed-junction* (or *fused-junction*) transistors. In sketch form, their structure is shown in Fig. 6-17. It is essentially an *n*-type semiconductor wafer, with two dots containing *p*-type impurities fused, or alloyed, into the opposite sides of the wafer to provide an emitter and a collector junction. Since the collector junction is to dissipate more power, it is usually made larger in size than the emitter junction. Alternatively, a symmetrical transistor can be fabricated in which the two junctions are the same in size.

Leads are then made to the emitter and collector, and the base connection is the shape of a ring as a way of reducing the series base resistance. The transistor is packaged in a sealed metal case through which the leads are passed inside glass insulators. In many transistors, one of the leads (the base or collector lead) is connected to the case.

In alloyed-junction transistors the base cannot be made very narrow, and so they are intended for operation at only low and medium frequencies. When a narrow base is produced by the alloying process, its width differs from one place to another. In order to avoid a punch-through, the voltage across the collector junction has to be lowered, and this reduces the power rating of the transistor.

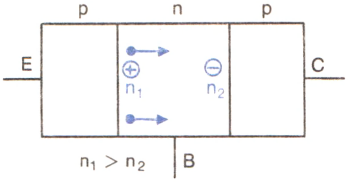


Fig. 6-16
Structure of a drift transistor

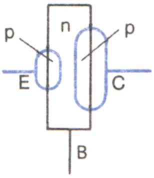


Fig. 6-17
Structure of an alloyed transistor

In high-power alloyed-junction transistors, the junctions have a larger surface area because they are made in the form of strips or rings. For better heat withdrawal, the collector of such transistors is welded to the case whose bottom is a substantial copper block (a *heat sink*). Because of this, the collector lead in high-power transistors is usually connected to the case.

Alloyed-junction transistors are available in power ratings from 10 mW to tens of watts. One of their advantages is that the collector and emitter junctions can withstand a reverse voltage of as high as 50-70 V in the case of germanium and 70-150 V in the case of silicon. This is an important advantage for high-power and switching transistors. Owing to the low emitter, base and collector resistances, alloyed-junction transistors are able to handle heavy currents when used in the switching mode. However, frequency limit, f_α, cannot be made higher than 20 MHz. Another limitation of alloyed-junction transistors is a marked spread in parameters and characteristics among transistors of the same type.

Drift transistors are made for operation at current-gain cutoff frequencies which are tens of times higher than they are for alloyed transistors. They owe this advantage above all to a reduction in the transit time of carriers through the base. As a rule, drift transistors are fabricated by the diffusion method which permits a very thin base to be made. Since a graded collector junction is produced, its capacitance is a small fraction of what it is in alloyed-junction transistors. Owing to a very narrow base, the alpha and beta current gains are appreciably greater than they are in alloyed-junction transistors. Importantly, the diffusion method makes it possible to fabricate transistors to closer tolerances and with a smaller spread in parameters and characteristics.

Alloyed-diffused (or *diffused-alloyed*) *transistors* differ in that they have an alloyed emitter junction and a diffused base region and a diffused collector junction. Many of the Soviet-made transistors are fabricated by the alloying-diffusion process. As an example, Fig. 6-18*a* shows a form of an alloyed-diffused germanium *p-n-p* transistor. It is basically a *p*-type germanium wafer used as the collector, in which two depressions are made where a thin base region is produced by diffusing a donor impurity, such as antimony. The base region and the wafer form between them a collector junction. The *p*-type emitter region is produced by fusing a dot of an alloy containing an acceptor impurity, such as indium, into the base region. The base lead is likewise produced by fusing a dot of an alloy containing antimony. In this design, it is usual to connect the collector to the case. A similar method can be used to make *n-p-n* transistors and silicon transistors. Alloyed-diffused transistors can operate at frequencies as high as hundreds of megahertz, but at low power levels (100-150 mW at most). They have a thin emitter junction, and so it can stand up only to low reverse voltages.

Conversion transistors are interesting in that they can be made with a thin base region of a large surface area so essential for high-power, high-frequency transistors. In conversion transistors, a diffused emitter junction is formed owing to the reverse diffusion of an impurity from the semiconductor into the metal of the emitter electrode. This purpose is served by a germanium wafer (the source material) which simultaneously contains donor and acceptor impurities. The acceptor impurity may be copper which, when the emitter alloy is fused into the wafer, diffuses eagerly from the germanium into the emitter. As a result, the concentration of the acceptor impurity in the germanium layer adjacent to the emitter is suddenly reduced and an *n*-type base region is thus produced. This reversal of conduction is called *conversion.*

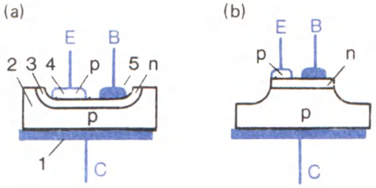


Fig. 6-18

Structure of (*a*) an alloyed-diffused transistor and (*b*) mesa transistor: (*1*) collector substrate; (*2*) collector (source material); (*3*) diffused base layer; (*4*) emitter alloy; (*5*) base lead alloy

Conversion transistors have a relatively low value of C_C and are able to operate at relatively high voltages across the collector junction. They show a good stability and a relatively small spread in parameters and are convenient to make. Unfortunately, they permit a very low maximum reverse voltage across the emitter junction.

Mesa transistors. This type of transistor is produced as has been explained with regard to diodes in Sec. 3-8. Owing to the production method used, a great number of mesa transistors can be fabricated at a time from a single wafer of the source semiconductor, so that the spread in parameters between the individual devices is greatly reduced. As the first step, a base layer a few micrometers thick is produced by the diffusion process on the surface of a germanium chip which is to be used as the collector. For each transistor, a tiny pellet of an appropriate impurity is fused into the base layer in order to produce an emitter region and the base terminal. Then, a masking compound is applied to small areas around the base and emitter and the remaining material of the chip is etched away. When the etching step is complete, the base and emitter regions appear as plateaus, or mesas, above the collector region. The wafer is then diced into individual transistors. In sketch form, the structure of a separated transistor is shown in Fig. 6-18*b*. As an example, it shows a *p-n-p* germanium transistor.

Mesa transistors have low values of junction capacitances (their C_C is less than 1 or 2 pF), a low value of r_B, and are able to operate at frequencies up to several megahertz. It is also convenient that heat can readily be drawn away from the collector because it has a terminal contact of a fairly large surface area.

The best among diffused transistors are *planar transistors.* In them the *p-n* junctions are formed by the diffusion of impurities through an opening in a protective coat applied to the surface of a semiconductor so that the emitter, base and collector regions come all to the same plane surface, that is, terminate in the same geometrical plane – hence the name 'planar'. Planar transistors are usually fabricated from silicon because the silicon oxide film on its surface can serve as an effective masking material. The silicon wafer covered by an oxide film forms the collector region. The area designed to serve as the base region is stripped off its oxide film by etching, and the base region is produced by the diffusion process. Then all of the surface is again given an oxide coat (this operation is known as *passivating*), after which the etching and diffusion steps are repeated to produce an emitter region which is situated in the middle of the base region. Following that, terminals are deposited in the form of metal layers through the mask. The structure of a planar transistor is shown in Fig. 6-19. Planar transistors have shown a consistently good performance and have found many applications. They are convenient to fabricate and can be made for different power ratings and high current-gain cutoff frequencies. It may be added that the planar technology is also used to fabricate transistor and diode elements in integrated circuits.

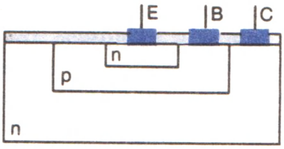


Fig. 6-19
Structure of a planar transistor

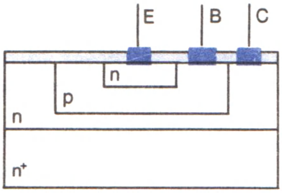


Fig. 6-20
Structure of a planar-epitaxial transistor

Epitaxial-planar (or *planar-epitaxial*) *transistors* are a further step forward in the planar technology. Conventional planar transistors have a collector region of a relatively high resistance which is a disadvantage. For example, when such a transistor is operating as a switch in the saturated region, its saturation resistance R_{sat} is very high. It could have been reduced by using a collector material of a lower resistivity, but this would have led to a higher value of C_C and a lower breakdown voltage of the collector junction. These limitations are all nonexistent in epitaxial transistors in which a layer of a higher resistivity is sandwiched between the base and the low-resistance collector. In the fabrication of such transistors, a thin layer of, say, *n*-type semiconductor which has a high resistivity is deposited onto a low-resistivity collector substrate of the same type of conduction, following which the base and emitter regions are produced by the planar technology (Fig. 6-20).

The process by which a layer having the same structure as the parent wafer but a different resistivity is deposited is called *epitaxial growth.* The resulting structure, designated as n^+-*n*, is part of the collector. The " + " sign labels the region which has a higher impurity concentration, that is, a higher conductivity.

The epitaxial-planar transistor we have taken as an example combines the low resistance of the collector with a low collector-junction capacitance C_C and a relatively high $V_{CB\,max}$. Epitaxial technology is widely used in the fabrication of integrated circuits.

There are other special transistor types which have not yet found a very broad field of application. They include, for example, *p-n-i-p* transistors in which the base has a relatively low-resistive *n*-type layer from which a terminal is made and also an additional *i*-type layer of a higher resistivity. The low-resistivity layer of the base brings down r_B, and its high-resistivity layer reduces C_C and raises $V_{CB\,max}$. Similar properties are found in *n-p-i-n* transistors.

Of special interest are *avalanche transit-time transistors* which utilize the avalanche multiplication of carriers, that is, they operate at V_{CB} exceeding the value considered normal for operation in the amplification mode. Under certain

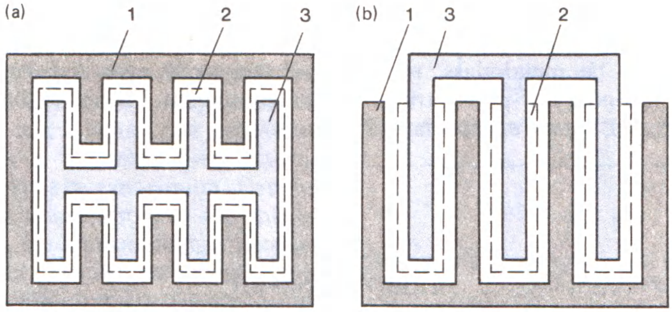


Fig. 6-21
Electrode structure of high-power microwave transistors: (*a*) interdigitated; (*b*) overlay. (*1*) Base lead; (*2*) emitter region; (*3*) emitter lead

conditions, an avalanche transit-time transistor will have a negative input resistance and an alpha current gain of more than unity. For this reason, they are ideally suited for use in pulse circuits to generate short pulses, and as switches.

Microwave transistors. It is attractive to use transistors for amplification at microwave frequencies because they need lower supply voltages than other semiconductor amplifying devices and tubes. However, their fabrication poses a number of engineering problems. At this writing, both low-power and high-power bipolar transistors fabricated from germanium, silicon and gallium arsenide are available commercially for operation at frequencies from units to several tens of gigahertz. The best results are obtained when microwave transistors are fabricated by planar technology. Notably, this technique is used to make microwave *n-p-n* transistors from silicon. High-power microwave transistors can deliver a pulse power of as much as 100 W at frequencies as high as 1 GHz, and 5 to 10 W at 4 or 5 GHz and higher.

Low-power transistors are extremely small in size. For example, a silicon wafer 40 mm in diameter can be used to form as many as 8 thousand transistors, each measuring 0.4 × × 0.4-mm square. As a rule, individual transistors are left unpackaged and are used in microcircuits.

Bipolar microwave transistors ordinarily use contacts (metallization) configured so that the emitter region consists of a number of narrow strips or small dots. This serves to reduce the series base resistance and the detrimental 'current crowding' effect – an outward bulging of electric flux lines at the edges. The contact (metallization) geometry is especially important for high-power transistors because the 'current crowding' effect is particularly strong at heavy currents. Several metallization geometries are currently in use. Two of them are shown in Fig. 6-21. The one in Fig. 6-21*a* is known as the *interdigitated structure.* Here the emitter takes the form of several long thin fingers (hence the name 'interdigitated' from the Latin *digit* for finger), and the emitter and base contacts (metallization) alternate. The other in Fig. 6-21*b* is known as the *overlay structure.* Here the emitter consists of many emitter dots (squares or circles). All the emitters are connected in parallel by the emitter metallization laid over a coat of protective oxide film. A large surface area is not needed for the junctions of low-power transistors, so they can be fabricated by the same principle but with a smaller number of fingers (Fig. 6-22) as little as 20-30 μm long and a few micrometers wide.

An important factor for the operation of a microwave transistor is the design of its case (package) and leads. They are chosen so as to minimize the effect of stray capacitances and inductances. Most commonly, cases have stripline leads, and those of higher-frequency devices, coaxial connections.

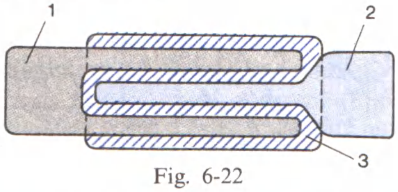


Fig. 6-22
Electrodes of a microwave transistor: (*1*) base; (*2*) emitter; (*3*) oxide film

Transistors are enclosed in cases, or packages, sealed against ingress of moisture and differing in design. They may be metal-glass, metal-ceramic, and plastic. Some low-power transistors are left unpackaged, but are encapsulated in a protective varnish or epoxy resin. In higher-power transistors, it is customary to connect the collector to the case, and the case is screwed to the equipment chassis for better heat abstraction.

Chapter Seven

Field-Effect Transistors and Thyristors

7-1 Field-Effect Transistors

Field-effect transistors (FETs for short), also known as *unipolar transistors* so as to tell them from bipolar devices, are extremely widely used semiconductor devices. The idea of using FETs was first advanced way back in 1952 by W. Shockley, one of the inventors of the bipolar transistor. A major advantage of FETs is a high input resistance which may be comparable with, or even greater than, that of vacuum tubes. At this writing, there is a growing trend for FETs to oust bipolar transistors in many applications.

The structure and connection of a *p-n* junction FET, or a JFET and its diagram symbol are shown in Fig. 7-1. It is essentially, say, an *n*-type semiconductor wafer which has at its opposite ends a pair of terminals (contacts or electrodes) for connection into the output (controlled) circuit of an amplifying stage. This circuit is energized from a supply voltage source E_2 and contains a load resistance, R_L. The output current flowing along a JFET is constituted by majority carriers which are electrons in our example. The input (control) circuit of the JFET is formed with the aid of a third electrode made to a region of an opposite type of conduction which is a *p*-type region in our example. The input supply voltage source E_1 produces a reverse voltage across the only *p-n* junction of the JFET. No forward voltage is applied to the *p-n* junction because the input resistance would have been very small. The signal source (or *signal generator*, *SG*) is connected in the input circuit.

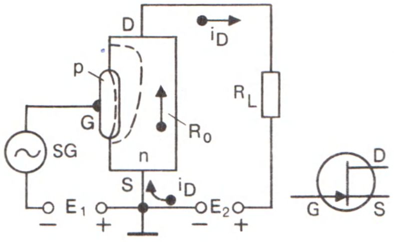


Fig. 7-1

Connection of an *n*-channel JFET in a circuit and its graphical symbol

Physically, a JFET operates as follows. When the input voltage is varied, a change occurs in the reverse voltage across the *p-n* junction, and this varies the width of the depletion (barrier) layer bounded by the dashed lines in Fig. 7-1. Accordingly, a change occurs in the cross-sectional area of the region through which the stream of majority carriers, that is, the output current, passes. This region is called the *channel.*

The electrode emitting majority carriers into the channel is called the *source* (*S*) and the electrode collecting them is called the *drain* (*D*). Obviously, the source and drain of a JFET are analogous to the cathode and anode of a vacuum tube, respectively. The control electrode intended to vary the cross-sectional area of the conducting channel is called the *gate* (*G*) and is similar to a certain extent to the grid of a vacuum triode or the base of a bipolar transistor. Of course, different physical principles underly the operation of the gate and the base.

As the gate voltage v_{GS} is increased, so does the width of the depletion layer in the *p-n* junction, but the cross-sectional area of the conducting channel is reduced. In consequence, its d.c. resistance R_0 increases and the drain current i_D decreases. At some value of the source-to-gate voltage, called the *pinch-off vol-*

tage and symbolized as v_{GSP},* the cross-sectional area of the channel reduces to zero and i_D falls to an extremely small value. As a result, the JFET is turned off. Conversely, at v_{GS} equal to zero, the channel has a maximum cross-sectional area, R_0 is a minimum (say, several hundred ohms), and i_D is a maximum. For the input voltage to be most effective in controlling the output current, the semiconductor material in which the channel is formed must have a high resistivity, that is, a low impurity concentration. Then, a wide depletion layer will be formed there. Also, the initial width of the channel (at $v_{GS} = 0$) must be sufficiently small. As a rule, it does not exceed a few micrometres. In the circumstances the pinch-off voltage will be several volts.

Since the potential builds up on moving along the channel towards the drain, the reverse voltage across the *p-n* junction is greater closer to the drain, and a wider depletion layer is formed there.

JFETs can be fabricated by the alloying method and the diffusion method. JFETs fabricated by the diffusion method show a better performance. Figure 7-2 illustrates the structure of a diffused JFET fabricated by epitaxial-planar technology. Devices with *n*-type source and drain regions and hence an *n*-type channel are referred to as *n*-channel JFETs and those with *p*-type regions as *p*-channel JFETs. The example shows a *p*-channel JFET (of course, an *n*-channel JFET could have been shown instead). The source and drain regions are usually made to have an enhanced p^+-type conduction so as to reduce the useless voltage drop across and the loss of power in those regions. This arrangement provides for an increase in the width of the depletion layer mainly towards the channel, thereby augmenting the control effect of the gate. The transistor chip (the substrate) is an *n*-type region from which a lead is often made. Then the substrate can be used as an additional gate. By applying, say, a constant voltage to it, it is possible to set the initial channel width.

Alloyed JFETs are low-frequency devices while diffused JFETs are able to operate at frequencies of tens or even hundreds of megahertz. It is to be stressed that the majority carriers are caused to move through the channel by an accelerating field at a very high velocity, so the frequency limit is determined by the inherent capacitances of a JFET and not by the transit time effects.

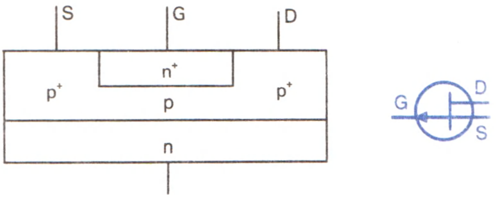


Fig. 7-2
Structure and graphical symbol of a planar-epitaxial *p*-channel JFET

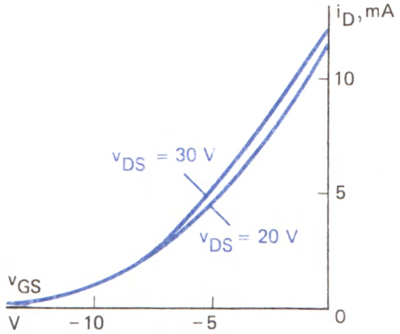


Fig. 7-3
Drain-gate characteristics of an *n*-channel JFET

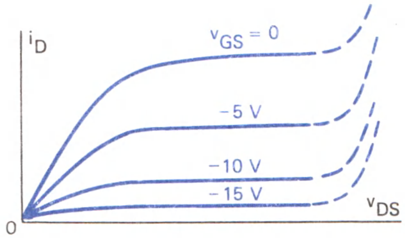


Fig. 7-4
Output (drain) characteristics of an *n*-channel JFET

The control action of the gate is illustrated by the *transfer*, or *drain-to-gate*, *characteristics*, $i_D = f(v_{GS})$ with v_{DS} held constant (Fig. 7-3). Unfortunately, these characteristics are not convenient for use in calculations, and resort is usually made to the *output* (or *drain*) *characteristics*.

As is seen from Fig. 7-4, the output (or drain)

* Some authors designate it as V_P.—*Translator's note.*

characteristics of a JFET relate the drain current, i_D, to the drain-to-source voltage, v_{DS}, that is, $i_D = f(v_{DS})$ with v_{GS} held constant. They show that with an increase in v_{DS} at first i_D rises at a fairly high rate, then its rise slows down, and finally ceases completely – the device reaches a state not unlike saturation. The point is that a rise in v_{DS} should have caused the drain current to rise as well, but since the reverse voltage across *p-n* junction also rises, the depletion layer widens, the channel's width is decreased (that is, its resistance goes up), and i_D is forced to fall. Thus, the current is exposed to two opposing effects at the same time so it remains practically constant.

When a high negative voltage is applied to the gate, i_D is reduced, and the characteristic is shifted downwards.

In the long run, a continuous rise in drain voltage leads to an electric breakdown of the *p-n* junction, and the drain current builds up cumulatively; this condition is shown in the figure by the dashed lines. The breakdown voltage is one of the absolute maximum ratings of JFETs.

As a rule, JFETs are operated within the quietly sloping portion of their characteristics, that is, within the region which is called, somewhat misleadingly, the *saturation region*. The voltage at which a JFET enters this region is referred to as the *saturation voltage*, and the pinch-off voltage is sometimes called the *cutoff voltage*.

It is to be noted that *p*-channel JFETs use supply voltages whose polarities are opposite to those shown in Figs. 7-1, 7-3, and 7-4 for *n*-channel JFETs.

Several parameters are used to specify the performance of JFETs. The most important one is the *mutual conductance* or *transconductance* g_m similar to the y_{21} parameter of bipolar transistors.* It is given by

$$g_m = y_{21} = \Delta i_D / \Delta v_{GS} \text{ with } v_{DS} = \text{const} \qquad (7\text{-}1)$$

and may take on values up to several milliamperes per volt.

The magnitude of transconductance is a measure of the control action of the gate. For example, $g_m = 3$ mA/V indicates that a change of 1 V in gate voltage will produce a change of 3 mA in drain current.

The second most important parameter is the *output resistance* r_{ds} (or r_d) similar to the $1/y_{22}$ parameter of bipolar transistors. This parameter gives the a.c. resistance of a JFET as measured between its drain and source (the channel resistance). It is given by

$$r_{ds} = 1/y_{22} = \Delta v_{DS} / \Delta i_D \text{ with } v_{GS} = \text{const} \qquad (7\text{-}2)$$

Within the quietly sloping portion of the output characteristics r_{ds} may be several hundred kilohms which is many times the d.c. resistance of a JFET, R_0.

Sometimes, a third parameter, called the *amplification factor* μ, is used. It shows how many times stronger a change in gate voltage affects drain current than a change in drain voltage. The amplification factor is given by

$$\mu = -\Delta v_{DS} / \Delta v_{GS} \text{ with } i_D = \text{const} \qquad (7\text{-}3)$$

In other words, it is the ratio between the incremental changes Δv_{DS} and Δv_{GS} which balance each other in their effect on i_D with the result that the drain current remains unchanged. For this to happen, the incremental changes in v_{DS} and v_{GS} must occur in opposite senses (they must take on different signs). This is the reason why there is a "–" sign on the right-hand side of Eq. (7-3). Alternatively, we may take the absolute value of the right-hand side. The amplification factor is connected to g_m and r_{ds} by a simple relation of the form

$$\mu = g_m r_{ds} \qquad (7\text{-}4)$$

Within the quietly sloping portions of the output characteristics μ may run into several hundreds or even thousands. At the start of the curves where they run steeply upwards (at low values of v_{DS}), all the three parameters take on small values. For a given set of operating conditions (a given mode of operation), g_m and r_{ds} can be derived from the output characteristics by the two-point method such as used to construct the load line for bipolar transistors (see Chap. 5), and μ can be found by Eq. (7-4).

The *input resistance* of a JFET is found in the usual way

$$r_{in} = \Delta v_{GS} / \Delta i_G \text{ with } v_{DS} = \text{const} \qquad (7\text{-}5)$$

Because i_G is a reverse current for the *p-n* junction and is therefore very small, r_{in} is units or tens of megohms. A JFET also has an *input capacitance* between its gate and source, C_{GS}, which is the barrier capacitance of the *p-n* junction and has a value of several picofarads

* More accurately it is termed the *common-source forward transconductance* and is often given the symbol g_{fs}.– *Translator's note.*

for diffused JFETs and tens of picofarads for alloyed JFETs. The *transfer capacitance* (that is, one from gate to drain), C_{GD}, has a lower value, the smallest capacitance being that between source and drain, C_{SD}, or the *output capacitance* of a JFET.

Similarly to bipolar transistors, a JFET may be connected in any one of three basic circuit configurations. Figure 7-1 shows the most commonly used *common-source* (CS) connection which is similar to the CE configuration. A CS stage yields a very high current and power gain and inverts the phase of voltage in the amplification mode. Because R_L usually is only a very small fraction of r_{ds}, the stage voltage gain may approximately be written as

$$k \approx g_m R_L \tag{7-6}$$

which is similar to Eq. (6-4) applicable to the CE circuit.

Figure 7-5 shows an equivalent circuit of a JFET with the CS connection. Since r_{in} is very high, it may be neglected. At low frequencies, we may also neglect all capacitances in many cases. The equivalent constant-current generator (the Norton equivalent) $g_m V_{m\ in}$ reflects the amplification supplied by the JFET, and r_{ds} is the a. c. resistance of the channel, that is, the output resistance. The signal source (or signal generator) is connected to the input terminals, and the load to the output terminals.

Practical amplifying stages ordinarily use a single supply voltage source E_2, as shown in Fig. 7-6 for *n*-channel JFETs. In order to derive the direct reverse voltage to be applied to the control *p-n* junction, the source lead contains a resistor R_S shunted by a capacitor C_S. The direct drain current I_{D0} produces across R_S a voltage drop

$$V_{GS0} = I_{D0} R_S$$

which is applied via the signal generator *SG* to the *p-n* junction. The value of R_S is found by the equation

$$R_S = V_{GS0}/I_{D0}$$

The values of V_{GS0} and I_{D0} required for a given set of operating conditions (or mode of operation) can be deduced from the output characteristics. The capacitor C_S provides a path for the alternating component of drain current. Its value must be such that its reactance at the lowest frequency, f_l, could be a small fraction of R_L. Then a small a. c. voltage drop will be produced across C_S. If C_S is not provided or its value has been chosen too small, a substantial a. c. voltage will appear across R_S. It will be applied to the transistor input in anti-phase with v_{in} (negative feedback), so that the resultant a. c. voltage at the transistor input will be reduced, and the amount of amplification will be smaller.

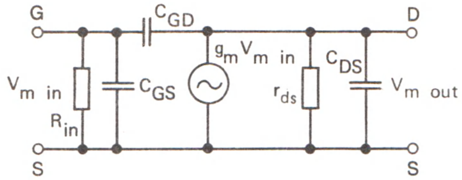


Fig. 7-5
Equivalent circuit of a JFET

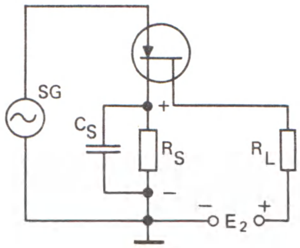


Fig. 7-6
JFET drawing power from a single supply

It is to be noted that this kind of negative feedback is sometimes deliberately used to improve the performance of an amplifier (by reducing distortion and improving the stability of its gain).

The circuit of Fig. 7-6, often called a self-biased circuit because V_{GS0} applied to the *p-n* junction is derived as a self-bias voltage, cannot be used where the JFET has to be turned off. To demonstrate, the bias voltage V_{GS} is produced by the flow of drain current I_{D0}, but this current is zero when the JFET is turned off. If there is a need for the JFET to be turned off in the absence of input voltage v_{in}, resort is made to the circuit of Fig. 7-7. Here, E_2 is applied to a divider $R_1 R_2$, and the direct voltage across R_1 is utilized as the cutoff bias voltage V_{GS0}. The value of R_1 is given by

$$R_1 = V_{GS0}/I_{dv}$$

where I_{dv} is the current through the divider; it is chosen relatively small so as to minimize the loss of power supplied by the E_2 source. On the other hand, I_{dv} must be many times I_{D0} which appears when V_{in} is applied. The capacitor C does the same job as in the previous scheme.

Sometimes the signal voltage supplied by the signal generator *SG* contains a direct component which ought not to reach the transistor input. To achieve this, the signal voltage is fed via a d. c. blocking capacitor C_b (Fig. 7-8), and the bias voltage V_{GS0} is applied via a resistor R_G which must have a high value so as to avoid a reduction in the input resistance of the stage.

Figure 7-9 shows an *n*-channel JFET connected in a common-gate (CG) circuit and a common-drain (CD) circuit. The CG configuration is analogous to the CB configuration. It does not produce current amplification, and so its power gain is a small fraction of that supplied by the CS circuit. Its input resistance is low because the input current is the drain current. When used as an amplifier, this scheme does not invert the phase of output voltage.

A common-drain (CD) stage (Fig. 7-9*b*) is similar to an emitter follower and may be called a *source follower*. The stage voltage gain is unity very nearly. The output voltage follows the input voltage in both phase and magnitude. This configuration has a relatively low output resistance and an increased input resistance. Also, it has a smaller input capacitance so that the input resistance is raised at high frequencies.

In addition to a high input resistance, JFETs have a number of other advantages over bipolar transistors. Because i_D in JFETs is due to the motion of majority carriers whose concentration is mainly governed by the impurity concentration and therefore is only slightly dependent on temperature, JFETs are less sensitive to temperature variations. They are able to operate over a wider range of temperature values. A rise in temperature only leads to a marked increase in the gate current (which is constituted by the minority carriers). Still, it remains sufficiently small, and the input resistance retains a high value. JFETs are less noisy and stand up better to ionizing radiations. In terms of radiation stability, JFETs come very closely to vacuum tubes. A disadvantage of many JFETs is a relatively small transconductance.

As a rule, JFETs are fabricated from silicon because the gate current, which is the reverse current for the *p-n* junction, is then only a small fraction of its value for germanium. At 20°C, the direct gate current may be as low as 1 nA, that is, 10^{-9}A.

The more recently developed and probably more important devices are *insulated-gate FETs* or *IGFETs*. In the most popular type which uses a metal gate, this metal gate is insulated from the semiconductor channel by a thin layer of oxide. This device is termed a metal on silicon dioxide

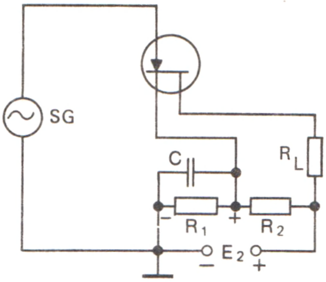


Fig. 7-7

JFET supply with provision for cut-off

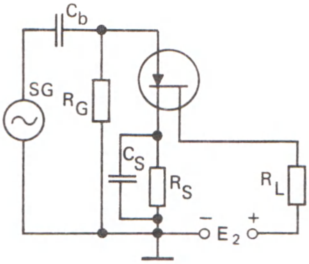


Fig. 7-8

Input voltage applied via a d. c. blocking capacitor

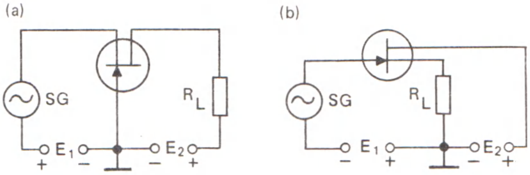


Fig. 7-9

JFET connected in (*a*) the common-gate (CG) configuration and (*b*) the common-drain (CD) configuration type or the *metal-oxide-semiconductor field-effect transistor*, or MOSFET.

The structure of and a graphical symbol for the MOSFET are shown in Fig. 7-10. The substrate is a high-resistance *p*-type silicon bar in which two low-resistance n^+-type regions are produced. One of them is the source, and the other is the drain of the device. Each region has a lead for connection to an external circuit. The source and the drain are separated by an *n*-type layer where a conducting channel is formed. Its length from source to drain is usually several micrometres, and its width is hundreds of micrometres or even greater, depending on the current that a given MOSFET is to handle in service. The SiO_2 layer (shown shaded) is 0.1 or 0.2 μm wide. On top of the oxide layer is the gate in the form of a thin metal film. The device substrate is usually connected to the source electrode, and its potential is taken as datum (zero), and so is the source potential. Sometimes, a separate lead is made to the chip, and such a device operates as described below.

Depending on the surface treatment, an *n*-type conducting channel may be formed and exist, with a gate voltage of zero. If, at a gate voltage of zero, we apply a voltage between drain and source, a current constituted by electrons will be flowing through the channel. No current will be flowing through the substrate because one of the *p-n* junctions is reverse-biased. If we apply to the gate a voltage which is negative with respect to the source and, in consequence, with respect to the substrate, a transverse electric field will be set up in the channel, which will sweep conduction electrons out of the channel into the source and drain regions and into the substrate. The channel is thus depleted of electrons, its resistance is increased, and the drain current is reduced. The more negative is the gate voltage, the smaller this current. This is termed *depletion-mode operation*, and a MOSFET operating thus is referred to as a *depletion-type* MOSFET.

Conversely, if no channel exists with a gate voltage of zero, one can be formed by applying a positive gate voltage, thereby attracting electrons from the source and drain regions and from the substrate to the channel and enhancing the channel conductivity and the drain current. Quite aptly, this is called enhancement-mode operation, and a FET operating thus is called an *enhancement-type* MOSFET.

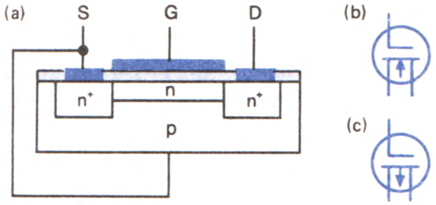


Fig. 7-10

IGFET: (*a*) structure, and graphic symbols for (*b*) *n*-channel IGFET and (*c*) *p*-channel IGFET

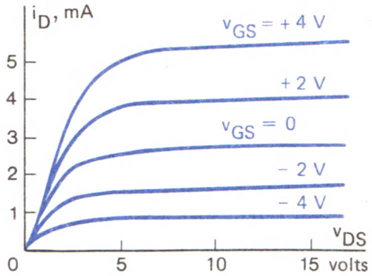


Fig. 7-11

Output characteristics of an *n*-channel depletion-type MOSFET

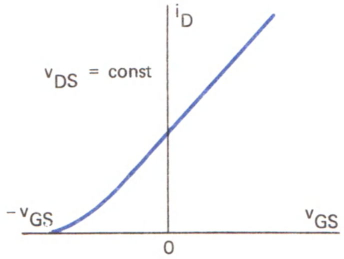


Fig. 7-12

Transfer characteristic of an *n*-channel depletion-type MOSFET

Likewise some FETs can operate in both the enhancement and depletion modes. This can clearly be seen from their output (drain) characteristics shown in Fig. 7-11 and the transfer characteristic of Fig. 7-12. As is seen, the output characteristics of MOSFETs are similar to those of JFETs. The explanation is that when

v_{DS} is raised, the device first obeys Ohm's law and the current rises approximately in proportion to the applied voltage. Then, at some value of v_{DS}, the channel's width begins to decrease, especially near the drain because of the rise in the reverse voltage which exists across the *p-n* junction between the channel and the substrate, the junction depleted of carriers grows wider, and the channel resistance is increased. Thus, the drain current is under the action of two opposing factors: a rise in v_{DS} tends to bring about an increase in the current in agreement with Ohm's law, but an increase in the channel resistance tends to cause the current to decrease. As a result, the current remains nearly constant until v_{DS} reaches a value causing a breakdown to the substrate.

If the substrate is an *n*-type material, the channel must be a *p*-type, and the polarity of voltages must be reversed. A graphical (diagram) symbol for a depletion-type MOSFET is shown in Fig. 7-10*c*.

In an enhancement-type MOSFET, a channel can only be formed and maintained when the device is in operation (Fig. 7-13), and then only when a definite value of gate voltage is applied in the correct polarity. In the absence of such a voltage, no channel exists, the source and drain of the n^+-type are separated solely by a *p*-type semiconductor substrate, and one of the p-n^+ junctions is reverse-biased. In this condition the resistance from source to drain is very high, and the MOSFET is turned off. If, however, we apply a positive voltage to the gate, the electric field set up at the gate will sweep conduction electrons out of the source and drain regions and the *p*-type substrate towards the gate. When the gate voltage exceeds some threshold value (several volts), the electron concentration in the surface layer grows so much that it exceeds the hole concentration, thereby giving rise to the *inversion of conduction* and to the *inversion layer* which serves as a thin (or narrow) *n*-type channel so that the MOSFET is turned on. The higher the positive gate voltage, the higher the channel conductivity and the greater the drain current. Thus, this type of MOSFET can only operate in the enhancement mode – this is corroborated by its output characteristics shown in Fig. 7-14 and the transfer characteristic in Fig. 7-15. If an *n*-type substrate is used, a *p*-type channel will be induced.

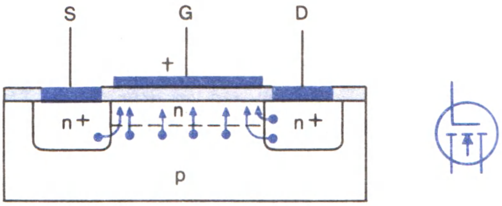


Fig. 7-13

N-channel enhancement-type MOSFET and its graphical symbol

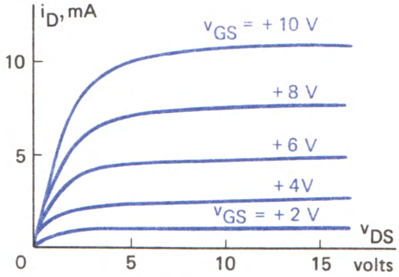


Fig. 7-14

Transfer characteristics of an *n*-channel enhancement-type MOSFET

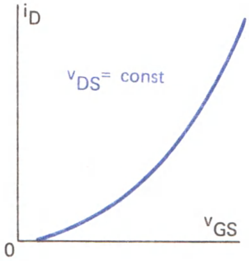


Fig. 7-15

Transfer characteristic of an *n*-channel enhancement-type MOSFET

The parameters of MOSFETs are analogous to those of JFETs. MOSFETs show a better performance in terms of temperature stability, noise level, radiation damage, and other properties mentioned in connection with JFETs. They offer a number of other advantages. Their d.c. input resistance at low frequency is the insulation resistance of the gate and is 10^{12}-10^{15} Ω. Importantly, the input resistance remains high with any polarity of gate voltage (in forward-biased JFETs the input resistance falls then to a very small value). The input capacitance may be less than 1 pF and the frequency limit runs into several hundred megahertz. High-power MOSFETs have been developed for which the transconductance is 10 mA/V and greater and which can operate at frequencies of several hundred megahertz. MOSFETs may be used in any of the three basic configurations examined earlier (CS, CG, and CD). It is to be noted that MOSFETs can be fabricated by planar-epitaxial technology with relative ease, and this simplifies the manufacture of integrated circuits. Enhancement-type MOSFETs are especially simple to make because only two doped regions have to be produced — the source and the drain.

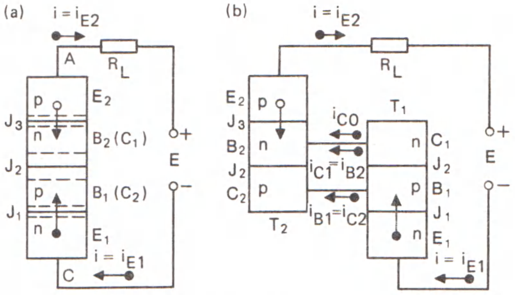


Fig. 7-16

Diode thyristor: (*a*) structure and (*b*) its equivalent circuit as a combination of two transistors

7-2 Thyristors

The term *thyristor* applies to a family of multilayer semiconductor devices that exhibit bistable (ON-OFF) switching action due to their inherent regenerative feedback. The term has its origin in the Greek *thyra* for door or entrance.

The structure of a *diode thyristor* of the *n-p-n-p* type is shown in Fig. 7-16*a*. As is seen, it has three *p-n* junctions, two of them (J_1 and J_3) being forward-biased and the third (J_2), which is sandwiched in between, being reverse-biased. The outer *p*-region is called the *anode*, and the outer *n*-region is called the *cathode*.

A thyristor may be modelled by an equivalent circuit as a combination of two transistors, T_1 and T_2, one being the *n-p-n* type and the other the *p-n-p* type, connected as shown in Fig. 7-16*b*. With this arrangement, J_1 and J_3 act as the emitter junctions of these two transistors, and J_2 operates as a collector junction in both transistors. The base region B_1 of transistor T_1 is at the same time the collector region C_2 of transistor T_2, and the base region B_2 of transistor T_2 is at the same time the collector region C_1 of transistor T_1. Accordingly, the collector current of T_1 (labelled i_{C1} in the diagram) is the base current for T_2 (labelled i_{B2}), while the collector current of T_2 (that is, i_{C2}) is the base current of T_1 (that is, i_{B1}). Experiments with a circuit made up of transistors as shown in Fig. 7-16 have proved that this setup is similar in its properties to a diode thyristor.

As a rule, thyristors are fabricated from silicon, with their emitter junctions alloyed and their collector junctions diffused. Planar technology is likewise used. The impurity concentration in the base (middle) region is substantially lower than it is in the emitter (outer) regions.

The physical processes that take place in a thyristor may be visualized as follows. If the device had only had one reverse-biased junction J_2, there would have been only a small reverse current flowing due to the movement of minority carriers across the junction. But, as will be recalled, a large collector current can be produced in a transistor, which is nevertheless a reverse current for the collector junction, if we inject a great number of minority carriers into the base from the emitter junction. The higher the forward voltage across the emitter junction, the greater the number of carriers reaching the collector junction and the higher the collector current. On the contrary, the voltage across the collector junction falls because when the current has a high value, the collector junction resistance decreases and there is an increase in the voltage drop across the load placed in the collector circuit. For example, in switching circuits the transistor is turned ON (that is, driven into the saturation region) by feeding an appropriate forward voltage to its emitter junction. Then the collector current reaches a maximum and the collector-to-base voltage falls to a few tenths of a volt.

A similar chain of events occurs in a thyristor. Minority carriers are injected across the forward-biased junctions J_1 and J_3 into the regions adjacent to the junction J_2 with the result that the resistance of J_2 is brought down.

The current-voltage characteristic of a thyristor, shown in Fig. 7-17, illustrates the events taking place in a thyristor when the applied voltage is raised. At first the current is small and rises slowly – this corresponds to portion OA of the characteristic. In this condition the thyristor may be deemed in the OFF state. The resistance of the collector junction J_2 is affected by two opposite factors. On the one hand, the rise in the reverse voltage across that junction tends to raise its resistance because the reverse voltage causes the majority carriers to move in opposite directions, that is, away from the boundary, and this implies that J_2 is progressively depleted of its majority carriers. On the other hand, the increase in the forward voltages across the emitter junctions J_1 and J_3 tends to step up the injection of carriers that arrive at J_2. In consequence, its carrier concentration is enhanced, with the result that its resistance increases, but does so progressively more slowly because the other factor produces an ever stronger effect.

Around point A, that is, at a certain definite voltage (tens or hundreds of volts), called the *forward breakover voltage* V_{fbo}, the two factors produce an equal but opposite effect, and balance each other. Now even a minute increase in the applied voltage gives the second factor an edge over the first, and the resistance of J_2 begins to decrease. Now the thyristor is driven to conduction (turned ON) in a cumulative fashion which can be explained as follows.

The current rises stepwise (portion AB of the characteristic) because an increase in voltage across J_1 and J_3 brings down the resistance of J_2 and the voltage across it. As a result, the voltages across J_1 and J_3 rise still more, and this leads in turn to a further increase in the current, a reduction in the resistance of J_2, and so on. This chain of events finally terminates in a condition which resembles the saturated state of a transistor – a heavy current at a low voltage (portion BC of the characteristic). The current in this state when the device is conducting (in the ON state) is mainly decided by the load resistance R_L connected in series with the device. Owing to the resultant heavy current, nearly all of the supply voltage is dropped across R_L.

In the ON state, large charges accumulate near junction J_2, and it is forward-biased – this condition is typical of the collector junction in the saturation region. Therefore, the total voltage across the thyristor is a sum of the three small forward voltages across the junctions and the four likewise small voltage drops across the p- and n-regions. Since each of these voltages is only a fraction of a volt, the total voltage across the thyristor in the ON state does not exceed a few volts and, in consequence, the thyristor presents a low resistance in this condition.

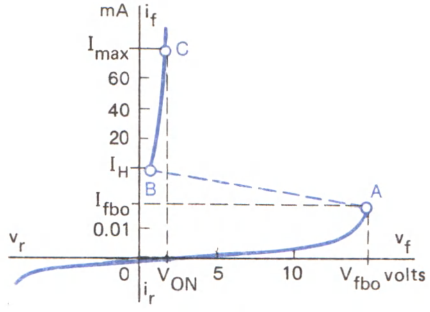


Fig. 7-17

Current-voltage characteristic of a diode thyristor

The stepwise transition of a thyristor from the OFF to the ON state may also be explained in a very simple way mathematically. From reference to the equivalent circuit in Fig. 7-16 it is seen that the thyristor current i is the first-emitter current i_{E1} or the second-emitter current i_{E2}. Alternatively, the thyristor current i may be regarded as a sum of two collector currents, i_{C1} and i_{C2}, respectively equal to $\alpha_1 i_{E1}$ and $\alpha_2 i_{E2}$, where α_1 and α_2 are the alpha current gains of transistors T_1 and T_2. Also, the thyristor current i includes the small reverse collector leakage current i_{C0}. Thus, we may write

$$i = \alpha_1 i_{E1} + \alpha_2 i_{E2} + i_{C0}$$

or (noting that $i_{E1} = i_{E2} = i$)

$$i = \alpha_1 i + \alpha_2 i + i_{C0} \tag{7-7}$$

On solving the above equation for i, we get

$$i = i_{C0}/[1 - (\alpha_1 + \alpha_2)] \tag{7-8}$$

Let us analyse Eq. (7-8). At small currents, α_1 and α_2 are substantially smaller than unity and their sum is likewise smaller than unity. Then, in accord with Eq. (7-8), the current i will be relatively small. However, α_1 and α_2 increase with increasing current, and this leads to a rise in i. At some value of current, called the *forward breakover current* I_{fbo}, the sum $\alpha_1 + \alpha_2$ becomes equal to unity, and the current i would build up without bound if it were not limited by the load resistance. This tendency of i to rise without bound points to a sudden stepwise increase in current, that is, to the fact that the thyristor turns ON.

Diode thyristors are specified in terms of the absolute maximum forward current $I_{f\ max}$ (point C in Fig. 7-17) at which a small ON-state voltage, v_{ON}, still exists across the thyristor. If we reduce the current flowing through the device to below some value called the *holding current*, I_H (point B), the current will drop suddenly and the voltage will rise as suddenly, indicating that the device has jumped back to the OFF state corresponding to the portion OA of the characteristic (called the *forward blocking region*). When a reverse voltage is applied to a thyristor, its current-voltage characteristic is the same as the reverse characteristic of a conventional semiconductor diode because junctions J_1 and J_2 are under reverse bias.

The other typical parameters of diode thyristors are the turn-on time t_{ON}, turn-off time t_{OFF}, total capacitance C_{tot}, maximum pulse forward current $I_{pulse\ max}$, and maximum reverse voltage $V_{r\ max}$. The turn-on time of thyristors is usually not over a few microseconds, and the turn-off time related to the recombination of carriers may run into tens of microseconds. Therefore thyristors are able to operate only at relatively low frequencies.

If a terminal lead is made to one of the base regions, the result will be a *triode thyristor*. By applying a forward voltage over this lead to the forward-biased junction, or the gate, we can vary the value of V_{fbo} at will. The greater the gate current I_G, the lower the value of V_{fbo}.

These principal properties of a triode thyristor are clearly demonstrated by its current-voltage characteristics shown in Fig. 7-18 for several values of gate current, I_G. The higher this current, the greater the number of carriers injected from the respective emitter to the middle collector junction, and the lower the value of voltage across the thyristor that is required for the device to start moving into the ON state. The highest value of V_{fbo} results when the gate current is zero and a triode thyristor becomes a diode thyristor. Conversely, when the gate current is high, the characteristic of a triode thyristor approaches the forward characteristic of a conventional semiconductor diode.

A simple connection diagram for a triode thyristor is shown in Fig. 7-19. Here, the graphical symbol is given for a thyristor with a third terminal lead made to its p-region. This is

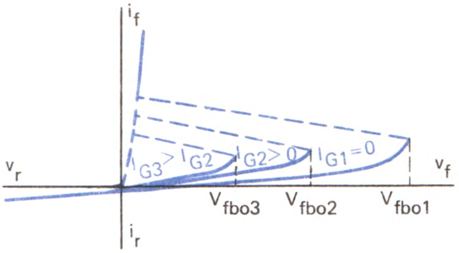


Fig. 7-18

Current-voltage characteristic of a triode thyristor for several values of control current

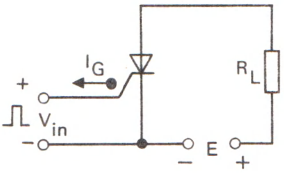


Fig. 7-19

Simple circuit using a thyristor with a lead from the p-region

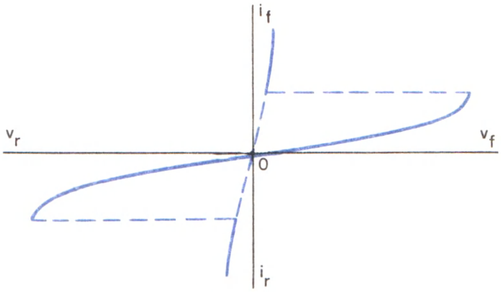


Fig. 7-20

Current-voltage characteristic of a symmetrical thyristor

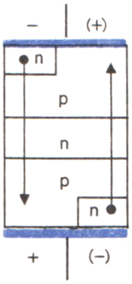


Fig. 7-21

Structure of a symmetrical thyristor

a *gate-triggered thyristor* because when a forward voltage pulse is fed over the gate lead to the emitter junction of the device, it is driven into the ON state, of course provided that the supply voltage E is sufficiently high.

Triode thyristors have the same parameters as their diode counterparts, with the addition of the quantities related to the gate circuit.

Ordinary triode thyristors cannot be turned off by control over the gate circuit. For a turn-off, the current through such a thyristor must be brought down below I_H. What are known as *gate turn-off thyristors* have however been developed and put to use, which can be turned off by applying a short reverse voltage pulse over the gate lead to the emitter junction. What are called *bidirectional thyristors* (also known as *symmetrical thyristors* or *symistors*) have also been developed. Their name refers to the passage of current through the device: bidirectional thyristors are capable of conduction in either direction as symbolized by the two-way arrows in their circuit symbols, and can be turned on by a voltage of either polarity (Fig. 7-20). They may have an *n-p-n-p-n* or a *p-n-p-n-p* structure.

In sketch form, the structure of a bidirectional thyristor is shown in Fig. 7-21. As is seen, when the polarity of the applied voltage is as shown by the "+" and "−" signs without brackets, the left-hand half of the device is operating (the direction in which electrons are moving is shown by arrows). When the polarity is reversed, as shown by the "+" and "−" signs in brackets, current is flowing the other way around, that is, through the right-hand half of the device. The job of a bidirectional thyristor can be done by two diode thyristors connected in parallel (Fig. 7-22). Controlled bidirectional thyristors have leads from the respective base regions.

Circuit symbols for various thyristors are shown in Fig. 7-23.

Triode thyristors are widely used in electronics, automatic control, and various industries. An example of how a triode (or a diode) thyristor can be used in a simple sawtooth voltage generator is shown in Fig. 7-24. A voltage source E charges a capacitor C at a relatively slow rate through a resistor R. So long as the voltage across the capacitor, v_C, remains low, the triode thyristor resides in the OFF state. When v_C becomes equal to the

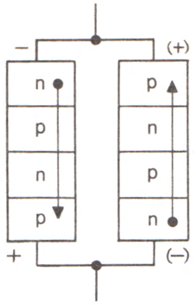


Fig. 7-22

Symmetrical thyristor replaced with two diode thyristors

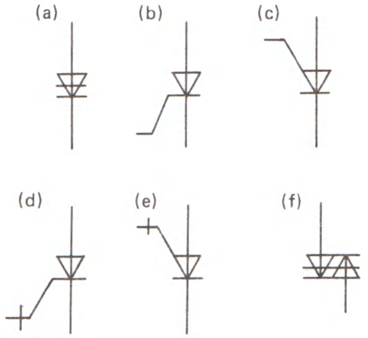


Fig. 7-23

Graphical (circuit) symbols of thyristors: (*a*) diode thyristor; (*b*) non-turnoff three-terminal thyristor with a lead from the *p*-region; (*c*) same, with a lead from the *n*-region; (*d*) gate turn-off three-terminal thyristor with a lead from the *p*-region; (*e*) same, with a lead from the *n*-region; (*f*) symmetrical thyristor

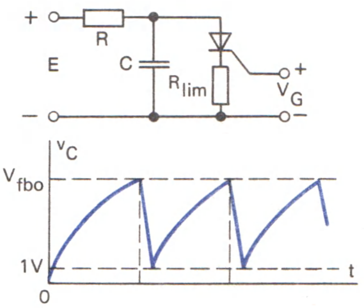


Fig. 7-24

Sawtooth voltage generator using a thyristor forward breakover voltage, V_{fbo}, the thyristor turns on and the capacitor discharges through the device at a very high rate because the resistance of a thyristor in the ON state is very small. At the end of the capacitor discharge, the current flowing through the thyristor falls to its holding value, I_H, and the device is turned off. After that, the capacitor is charged again, then it discharges through the thyristor, etc. A plot of the voltage generated across the capacitor is shown in Fig. 7-24. The purpose of R_{lim} is to prevent the current flowing through the resistor from exceeding its maximum rating. The greater the values of R and C, the longer the time it takes for the capacitor to charge, and the lower the frequency of the resultant voltage. Its amplitude is determined by the value of V_{fbo} and can be adjusted by varying the gate voltage V_G. As a rule, a second current-limiting resistor is placed in the gate circuit as well.

The sawtooth voltage produced by the generator we have just discussed may prove unsatisfactory for many applications because the voltage magnitude builds up exponentially. For a linearly rising voltage to be generated, the charging current of the capacitor must be held at a constant value. To this end, the resistor R may be replaced with a transistor connected in a common-base configuration. Then, as will be recalled, a change in v_{CB} will leave the collector current almost constant.

An interesting application for triode thyristors is in sinewave oscillators. In a sinewave oscillator, the thyristor operates as a switch to connect a supply source to the tank circuit at a desired rate. Owing to this arrangement, the tank circuit maintains undamped oscillations and the thyristor itself is controlled by the voltage taken from the tank circuit. Thyristor oscillators have a high efficiency because the loss of power in the thyristor itself is insignificant. Since, however, the turn-on and, especially, turn-off times are finite and entail a delay, such oscillators can only be operated at relatively low frequencies. Because thyristors are rated for heavy currents, thyristor oscillators may be built for substantially greater power values than transistor oscillators.

Chapter Eight

Miscellaneous Semiconductor Devices

8-1 The Tunnel Diode

In 1958, L. Esaki of Japan proposed a *p-n* junction diode that is made from germanium or gallium arsenide with an extremely high level of impurity concentration (or doping) on each side of the junction (of the order of 10^{19}-10^{20} cm^{-3}). This implies a very low resistivity – by a factor of several hundred or several thousand lower than it is in conventional *p-n* junction diodes. Such devices are said to be doped into degeneracy. The *p-n* junction in a degenerate semiconductor is by a factor of several tens thinner (10^{-6} cm) than it is in conventional diodes, while the potential barrier is twice as high. In ordinary semiconductor diodes, the height of the potential barrier is about half the forbidden band (or band gap); in the Esaki diode, now more usually called the tunnel diode, the figure is somewhat greater than the forbidden band width. Owing to a thin junction, the field intensity in it may be as high as 10^6 V cm^{-1} even when no external voltage is applied.

In a tunnel diode, as is the case with ordinary semiconductor diodes, carriers diffuse across the *p-n* junction and drift in the reverse direction under the influence of the field. Both processes, however, are of minor importance as compared with what has come to be known as the *tunnel effect* or *tunnelling*. It consists in that, in agreement with the laws of quantum physics, electrons have the chance of crossing the potential barrier without any change in their energy, even though they do not have sufficient kinetic energy to surmount it, provided the potential barrier is rather thin. This tunnelling of electrons with energies (in electron-volts) smaller than the barrier height can take place in both directions, provided there are unoccupied energy levels for the tunnelling electrons on the other

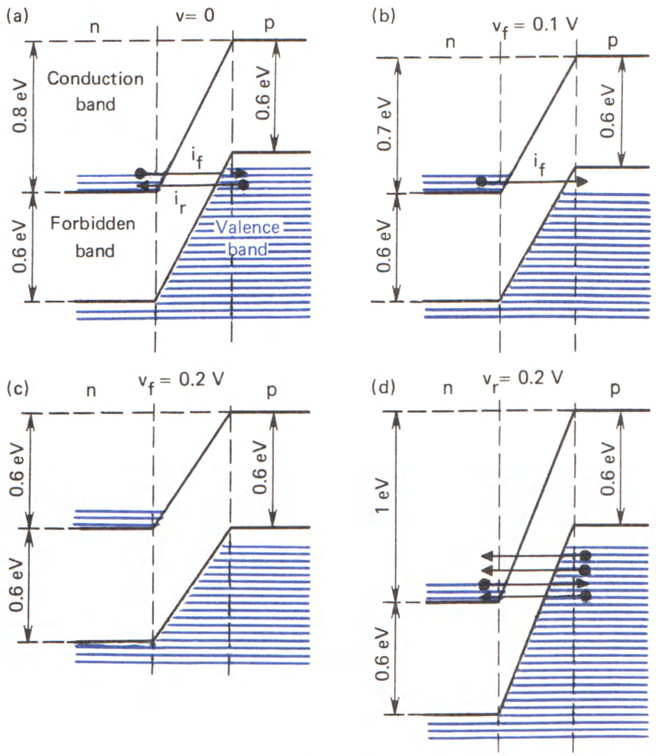


Fig. 8-1

Energy-band diagrams of the *p-n* junction in a tunnel diode for several values of applied voltage

side of the barrier. Classical physics rules out such an effect because it treats an electron as a material particle carrying a negative charge, but it is a reality in the microscopic world obeying the laws of quantum mechanics which treats an electron as a dual entity: it can behave both as a particle and as an electromagnetic wave. As an electromagnetic wave, it can cross the potential barrier, which is a region of an electric field, without interacting with the field.

What happens in a tunnel diode can conveniently be traced by reference to energy level diagrams showing the energy levels in the valence and conduction bands of *n*- and *p*-regions. Owing to the contact difference of potential arising at a *p-n* junction, the boundaries of all bands in one of the regions are shifted relative to the bands in the other region by an amount equal to the height of the potential barrier, expressed in electron-volts.

The energy level diagrams in Fig. 8-1 show how tunnel (or tunnelling) currents arise across the *p-n* junction of a tunnel diode. For simplicity, the diagram does not show the diffusion current and the conduction current. The diagram in Fig. 8-1*a* represents the case when no external voltage is applied. The height of the potential barrier is taken, for purposes of our example, equal to 0.8 eV, and the forbidden-band (band-gap) width is 0.6 eV. The horizontal lines in the conduction and valence bands represent the energy levels fully or partly filled by electrons. The unshaded portions in the valence and conduction bands represent the levels not occupied by electrons (empty states). As is seen, the conduction band of the *n*-type material and the valence band of the *p*-type material have filled energy levels corresponding to the same values of energy. Therefore, electrons are able to tunnel from the *n*- into *p*-region (the forward tunnel current i_f) and from the *p*- into the *n*-region (the reverse tunnel current i_r).

The two currents are equal in magnitude but opposite in polarity, so they cancel each other, and the net current is zero.

Figure 8-1*b* is an energy level diagram when a forward bias voltage of 0.1 V is applied to the device. Owing to this voltage, the height of the potential barrier is raised by 0.1 eV to 0.7 eV. This augments the tunnelling of electrons from the *n*- into the *p*-region because the valence band of the *p*-region has unfilled levels or states corresponding to the same values of energy as are possessed by the levels occupied by electrons in the conduction band of the *n*-region. On the other hand, no electrons can pass from the valence band of the *p*-region into the *n*-region because the levels filled by electrons in the valence band of the *p*-region correspond to the energy levels of the forbidden band in the *n*-region. There is no reverse tunnel current flowing, and the resultant tunnel current reaches a maximum. In the intermediate cases, such as when $v_f = 0.05$ V, both currents (forward and reverse) coexist, but the reverse current is lower in magnitude than the forward current. The net current will then be a forward current which is smaller in value than the maximum current which results when $v_f = 0.1$ V.

The case shown in Fig. 8-1*c* corresponds to $v_f = 0.2$ V when the height of the potential barrier is 0.6 eV. Here, too, no tunnelling can take place because the levels filled by electrons in one region correspond to the energy levels of the forbidden band in the other region. The tunnel current is zero. It will also be zero at a very large forward voltage.

It is important to remember that a rise in the forward voltage brings about an increase in the forward diffusion current of the diode. At v_f less than 0.2 V, the diffusion current is only a fraction of the tunnel current, while at v_f greater than 0.2 V the diffusion current rises to values typical of the forward current in a conventional diode.

Figure 8-1*d* illustrates the case where $v_r = 0.2$ V. The height of the potential barrier is then 1 eV, and there is a marked increase in the number of energy states filled by electrons in the valence band of the *p*-region and corresponding to empty states in the conduction band of the *n*-region. Therefore, there is an abrupt increase in the reverse tunnel current which is of the same order of magnitude as the current under forward bias.

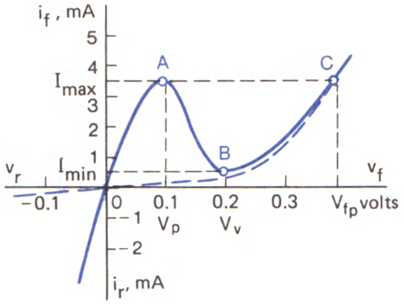


Fig. 8-2

Current-voltage characteristic of a tunnel diode

The current-voltage characteristic of a tunnel diode (Fig. 8-2) gives a further insight into the energy level diagrams used above. As is seen, at $v = 0$, current is zero. An increase in the forward voltage to 0.1 V leads to a rise in the forward tunnel current to a maximum (point *A*). The further increase in forward voltage to 0.2 V is accompanied by a decrease in the tunnel current. Therefore, the current at point *B* is a minimum, and the characteristic has a down-sloping portion, *AB*, corresponding to a negative resistance to alternating current

$$R_i = \Delta v/\Delta i < 0 \qquad (8\text{-}1)$$

Past that region, the current again rises owing to the forward diffusion component whose characteristic is shown in Fig. 8-2 by the dashed line. The reverse current is the same as the forward current, that is, many times the figure associated with conventional diodes.

The basic parameters of the tunnel diode are the *peak current* I_p (it is the value at which negative resistance and, in consequence, negative conductance first occur); the *valley current* I_v (it is the value at which negative resistance and negative conductance cease). The two corresponding voltages are the *peak voltage* V_p and the *valley voltage* V_v. The ratio I_p/I_v is termed the *peak-to-valley ratio*, and the difference $\Delta V = V_p - \mathrm{V}_v$ is termed the *voltage swing*. There is a second peak voltage, V_{fp}, on the up-sloping portion of the curve, corresponding to the peak current I_p (portion *BC*). Sometimes, the voltage swing is defined as the difference between V_{fp} and V_p. Currents in state-of-the-art tunnel diodes are several milliamperes, and voltages are a few tenths of a volt. The parameters of the tunnel diode also include its negative incremental resistance (usually of the order of several tens of ohms), the total capacitance of the diode (units or tens of picofarads), the *switching time* τ_{sw}, and the *maximum* or *critical frequency* f_{max}.

By connecting a tunnel diode in a circuit in a suitable manner, it is possible to use its negative resistance so as to balance the positive resistance (if its operating point is located within the portion *AB*) and to obtain an amplifier or an oscillator. For example, an ordinary resonant circuit always suffers a loss of power, and so it can only generate damped or decaying oscillations. This loss can however be made up for owing to the negative resistance of a tunnel diode and the resonant circuit will then generate undamped or sustained oscillations. A simple oscillator circuit using a tunnel diode is shown in Fig. 8-3.

Basically, the operation of such an oscillator can be explained as follows. When the supply voltage is turned on, free oscillations are initiated in the *LC* resonant circuit. Without a tunnel diode, they would have died out. Let *E* be chosen such that the diode is operating within the down-sloping portion of its characteristic and let during one half-cycle the resonant-circuit alternating voltage be of the polarity shown in the figure by the "+" and "−" signs without circles (the encircled "+" and "−" signs apply to d. c. voltages). The resonant-circuit voltage is fed to the diode so that it is reverse-biased. Therefore, the forward voltage across the diode is brought down. Since, however, the diode is operating within the down-sloping portion of its characteristic, its current rises so that an additional current pulse is produced, adding more energy to the resonant circuit. If this energy is large enough to make up for the power loss, the oscillations in the resonant circuit will not die out — they will be sustained.

Electrons tunnel through the potential barrier during an extremely short time: 10^{-12}-10^{-14} s or 10^{-3}-10^{-5} ns. Therefore, tunnel diodes do their job especially well at microwave frequencies. For example, they can generate and amplify signals at frequencies as high as tens or even hundreds of gigahertz. It is to be noted that the frequency limit of tunnel diodes is practically governed by the diode capacitance, its lead inductance, and its series resistance rather than by the inertia of the tunnel effect.

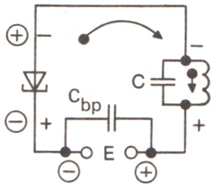


Fig. 8-3

Connection of a tunnel diode as an elementary oscillator

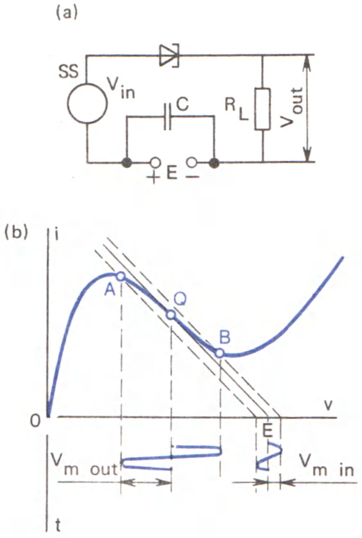


Fig. 8-4

(*a*) Tunnel diode as an amplifier and (*b*) plot explaining amplification action

The principle underlying signal amplification by the tunnel diode is illustrated in Fig. 8-4. For signal amplification to be possible, *E* and R_L must have certain definite values. R_L must be slightly smaller than the absolute value of the diode's negative resistance. Then, with no input voltage applied, the operating point *Q* can be positioned in the middle of the down-sloping portion of the characteristic (this point lies at the intersection of the load line and the diode characteristic). When an input voltage of amplitude $V_{m\ in}$ is applied to the diode, the load line will oscillate; in doing so, it will move parallel to itself. Its extreme positions are shown by dashed lines which locate the final points of the operating region *AB*. Projecting these points onto the voltage axis gives the peak value of output voltage, $V_{m\ out}$, which is substantially greater than the input voltage. A distinction of a tunnel-diode amplifier is the absence of separate input and output circuits, and this poses problems in setting up circuits with several amplifying stages. Tunnel-diode amplifiers are able to give a high gain at a low noise level and show a consistent performance.

Tunnel diodes can be used as fast-operating switches, with a switching time of around 10^{-9} s, that is, close to 1 ns, or even less. In the simplest case, the circuit of a tunnel diode operated as a switch is similar to that shown in Fig. 8-4, but the input voltage takes the form of pulses, and R_L must be somewhat greater than the absolute value of the diode's negative resistance. Operation of a tunnel diode in the switching (pulsed) mode is illustrated in Fig. 8-5. The supply voltage E is chosen such that in the absence of an input pulse the diode is operating at point A and the current is a maximum, I_p, that is, the diode is in the ON state (conducting). When a positive-going pulse of input voltage is applied, the forward voltage across the diode rises, and the diode jumps to operate at point B. The current falls to its valley value, I_v, which may be taken as the OFF state of the diode. If we adjust the value of E so that it corresponds to operation at point B, we will be able to move the diode to operation at point A by applying negative-going voltage pulses.

Tunnel diodes can well serve in microwave and pulse circuitry intended for high operating speeds. In addition to a very small inertia, tunnel diodes offer the advantage of high stability towards ionizing radiations. A low power drain from the supply source may also be regarded as an advantage of tunnel diodes in many cases. Unfortunately, they are not free from a serious drawback which consists in that they age rather rapidly – with time their parameters and characteristics may change to such an extent that the circuit in which the diodes are used might fail to operate properly. It may be hoped, however, that this limitation will be overcome some time in the future.

If the material used to fabricate a tunnel diode has an impurity concentration of about 10^{18} cm^{-3}, there will be practically no tunnel current flowing under forward bias, and the current-voltage characteristic will not have a negative-resistance (down-sloping) portion (Fig. 8-6). Under reverse bias, however, the tunnel current will be substantial as usual, and so such a diode will conduct current well in the reverse direction. Such devices have come to be known, quite appropriately, as *inverted diodes*; they can effectively operate as detectors at far higher frequencies than the usual diodes.

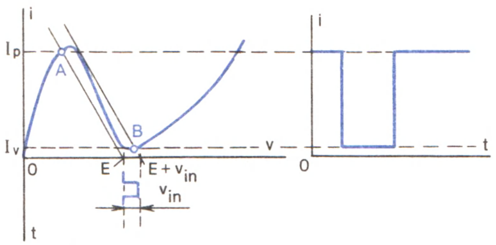


Fig. 8-5

Pulsed operation of a tunnel diode

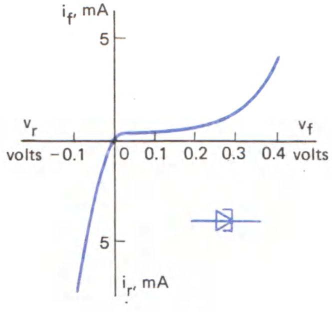


Fig. 8-6

Current-voltage characteristic and graphical (circuit) symbol of an inverted diode

All tunnel diodes are very small in size: several hundreds of them could be packed in a volume of 1 cubic inch. They can be enclosed in sealed cylindrical metal-glass cases 3 or 4 mm in diameter and about 2 mm high. Connection to an external circuit is by means of flexible ribbon leads. The mass of a typical enclosed tunnel diode is not over 0.15 g.

At this writing, work is under way on novel types of tunnel diodes, new semiconductor materials are being investigated for their fabrication, and the problem of ageing is being tackled.

8-2 Microwave Semiconductor Diodes

Microwave circuits widely use low-power point-contact semiconductor diodes. The materials used for their fabrication are germanium, silicon, and gallium arsenide with an enhanced donor or acceptor doping level owing to which the base has a low resistivity. This cuts down the lifetime of carriers and the storage time. Also the small surface area of the *p-n* junction results in the low value of junction capacitance. Exactly these features make such diodes idealy suited for use at microwave frequencies. With a base of low resistivity, however, the *p-n* junction is very thin, and its punch-through occurs even at a reverse voltage of as low as a few volts. In many cases, this is not a limitation because microwave diodes are mostly used under small-signal conditions. Still, diodes possessing a low breakdown voltage may readily fail at relatively small overvoltages, such as may be caused by a static charge.

As a rule, microwave semiconductor diodes are coaxial in structure (Fig. 8-7) so that they can conveniently be mated to coaxial lines or waveguides. The coaxial terminals of microwave diodes do away with the detrimental effects of lead capacitances and inductances. The type of diode shown in Fig. 8-7 is not the only one commercially available.

Uses to which *p-n* junction diodes can be put at microwave frequencies are many and diverse. Thus, there are *detector diodes* (also known as *video detectors*) which are employed in receivers and instruments over the entire microwave range. There are also *mixer diodes* used in microwave equipments for frequency conversion. As compared with vacuum-tube mixers, semiconductor units are advantageous in that the transit-time effect and the device capacitance are substantially smaller than they are in vacuum diodes. Also, semiconductor diode mixers have a lower inherent noise level.

Still another type of microwave diodes are *parametric diodes*. They are mostly used in low-noise parametric amplifiers. In these amplifiers, the diodes do the job of a nonlinear capacitor whose capacitance varies under the influence of the applied a. c. voltage. Parametric microwave amplifiers have a high gain coupled with a low noise level.

A further type is the *multiplier diode* which, as its name implies, is used for frequency multiplication. Since a semiconductor diode is a nonlinear device, it may sometimes be used as a *microwave modulator*.

Microwave circuits, notably those handling high power, can readily be switched by *switching diodes*. With them, electronic equipment can be made lighter in weight, smaller in size, more reliable, and more durable. Very little power is lost in switching microwave diodes, still their absolute maximum power rating is only a fraction of that achievable with some other switching devices.

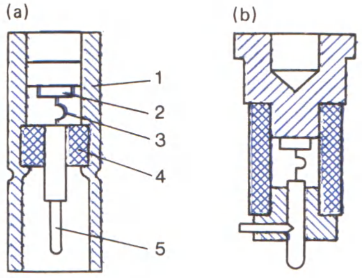


Fig. 8-7

Coaxial microwave diodes: (*1*) first electrode; (*2*) crystal; (*3*) needle; (*4*) insulator; (*5*) second electrode

The switching process consists in that the total resistance of the diode is made to change abruptly either by the input signal itself or by an additional direct control voltage applied to the diode in the required polarity. Several types of switching diodes are in use. As a rule, the difference between the forward and reverse resistances is utilized in *nonresonant switching diodes* which are fabricated to have a minimal capacitance and a minimal inductance. Therefore, such diodes are made without a case and terminal leads, and the capacitance of the *p-n* junction is balanced out by connecting an inductance to the device.

Resonant switching diodes operate as follows. When forward-biased, they are each a parallel resonant circuit made up of the case capacitance, the lead inductance, and the series (loss) resistance of the diode. At the resonant frequency, this circuit presents a very high impedance. Under reverse bias, however, the diode is re-configured into a series resonant circuit consisting of the lead inductance, the barrier (junction) capacitance, and the series (loss) resistance. In this condition, the diode presents a very low impedance at the resonant frequency. A resonant switching diode must have parameters providing for both parallel and series resonance at the desired frequency. Sometimes, additional inductive elements may be connected to the diode in order to achieve this objective.

Since point-contact diodes are able to switch only small power, one has to use junction diodes in cases involving several watts of power in continuous-wave (CW) operation. In pulse work, point-contact diodes are able to switch several kilowatts of power with a pulse duration of several microseconds. The switching time may then be not more than 10-20 ns.

Fast switching of microwave circuits that handle high power uses *p-i-n* junction diodes ordinarily made of silicon. These diodes have *p*- and *n*-regions of a relatively high conductivity (with a high doping level), separated by a broader *i*-type region so that the device has a low capacitance (Fig. 8-8). With no external voltage applied, the resistance of the *i*-region is high, and it grows still more under reverse bias because it causes a majority carrier depletion. Forward bias steps up the injection of holes from the *p*-region and of electrons from the *n*-region into the *i*-region. As a result, the resistance of the *i*-region and of the diode as a whole is abruptly reduced (by a factor of 10^3-10^4). In high-power *p-i-n* junction diodes the breakdown voltage for the *i*-region is several kilovolts, and so these diodes may be used to switch a pulse power of tens of kilowatts.

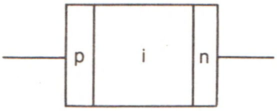


Fig. 8-8
Structure of a *p-i-n* junction diode

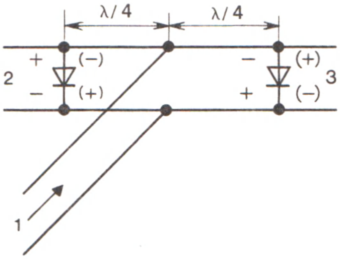


Fig. 8-9
Switching circuit using diodes

Quite a number of circuits using microwave switching diodes have been developed, each tailored to a particular task. As an example, Fig. 8-9 shows one such circuit. When the left-hand diode is forward-biased and the right-hand diode is reverse-biased (which is shown by the polarity signs without brackets), the left-hand diode presents a low resistance, and the right-hand diode, a high resistance. Therefore, the signal will practically be able to pass only from line *1* into line *3*, and the quarter-wavelength section of the left-hand line, nearly short-circuited by the diode, will act as a metal insulator. When the d.c. voltages are reversed (as shown by the polarity signs in brackets), the diodes will exchange their roles, and the signal will pass into line *2*.

A more recent trend with regard to switching and other functions in microwave circuits has been to use *Schottky-barrier diodes* (also called *Schottky* or *hot-carrier diodes*). The device is a form of a rectifying diode in which a junction is formed between a metal and a semiconductor – the *Schottky barrier*, thus called after W. Schottky of Germany who was the first to investigate what has come to be known as the Schottky effect (see Sec. 2-4). More specifically, a Schottky diode consists of a low-resistivity semiconductor substrate (which may be n^+-type silicon) with a high donor doping level, overlaid by a thin film of the same semiconductor material but having a high resistivity, in turn topped by a metal layer. A forward bias voltage is applied with its "+" side to the metal, and nearly all of the applied voltage is impressed on the high-resistivity film. Electrons in the film are accelerated to a very high velocity (they become 'hot carriers'), overcome the potential barrier, and enter the metal. However, no carrier storage occurs in the metal as it does in the base of a conventional diode. Therefore, Schottky diodes have a high operating speed which solely depends on the transit time of electrons through the high-resistivity film (less than 10^{-11} s) and on the barrier (or junction) capacitance which can be made very low owing to the small contact area. As a result, Schottky diodes are able to operate at frequencies as high as 15-20 GHz, and their switching time is a few tenths or even hundredths of a nanosecond. Their reverse current is very small.

8-3 Avalanche Diodes and Gunn Diodes

It has been recently discovered that *p-n* junction diodes operating in the reverse-bias direction and in the avalanche breakdown region are able to generate microwave oscillations and to perform signal amplification. The theory of these devices was first formulated by A. S. Tager and V. M. Wald-Perlov of the Soviet Union and stems from the negative resistance inherent in the two-carrier flow under space charge conditions. This negative resistance exists only at microwave frequencies and is nonexistent at lower frequencies.

Let a direct reverse voltage and some a. c. voltage be applied to an avalanche diode. During the positive half-cycles of reverse voltage (it is presumed that during these half-cycles the reverse voltage across the diode is rising) the diode is biased well into the avalanche breakdown region and the current is building up cumulatively. Owing to the inertia of processes in semiconductors, that is, the finite transit time of carriers crossing the *p-n* junction, this avalanche reaches its maximum with some delay from the positive half-cycle of a. c. voltage that has triggered it. Under the influence of the direct voltage, the avalanche keeps moving also during the subsequent negative half-cycles of voltage so that the current pulse associated with the avalanche is opposite in sign to the negative half-cycle of a. c. voltage. In other words, a negative resistance is now presented to a. c. When an avalanche diode is connected to a microwave resonant system, its negative resistance can be used to generate microwave oscillation or to perform amplification. At lower frequencies, the transit-time effect is negligible, and so is the delay of the current pulse with respect to the a. c. voltage, therefore, the negative incremental resistance is practically nonexistent. The *p-n* structure is not the only choice for avalanche diodes. For example, the Read diode has an n^+-p-i-p^+ structure.

In microwave oscillators, an avalanche diode is connected to a cavity resonator. In CW operation, such oscillators can deliver an output power of several watts at an efficiency of about 10%. In pulsed operation, their power output may be as high as several hundred watts at an efficiency of several tens of percent. Such oscillators can be tuned electronically (to within a few tenths of one per cent) by varying the applied d. c. voltage. Tuning over a wider range (tens of per cent) can be achieved by varying the natural frequency of the cavity resonator. When used as amplifiers, avalanche diodes suffer from a major disadvantage – a relatively high level of inherent noise.

Another example of semiconductor devices displaying a negative resistance at microwave frequencies is the *Gunn diode* which is based on the effect discovered by J. B. Gunn in 1963. This effect occurs when a large d. c. electric field is applied across a short sample of *n*-type gallium arsenide. At values above the threshold values (typically several thousand volts per cm), coherent microwave oscillations are generated. It took several years for Gunn himself and other investigators to study the effect in detail, to elucidate the physical processes it involves at high electric field intensities in semiconductors, and to develop practical devices capable of generating microwave oscillations.

The Gunn (or *transferred-electron*) diode is a device formed from a bar of low-resistivity *n*-type gallium arsenide (with no *p-n* junction produced) in which a very strong electric field is set up. The diode is connected to an external circuit by way of its two ohmic contacts one of which is called the anode and the other, the cathode. The reason for the use of gallium arsenide as the source material is that the diode must have two conduction bands. Studies of such semiconductors have shown that the two conduction bands differ in electron mobility. The electrons occupying the higher conduction band, that is, one corresponding to higher energy levels, have a lower mobility.

When no or a very weak electric field is applied, the electrons reside in the lower conduction band where they have a higher mobility and the material has a relatively high conductivity. When the applied voltage is raised, the current in the device first rises in accord with Ohm's law until a voltage is reached when the field intensity becomes sufficiently high for the greater proportion of electrons to be transferred to the upper conduction band where their mobility is reduced and the resistance of the material there is abruptly increased. The current falls, and a down-sloping portion appears in the current-voltage characteristic, indicative of a negative differential resistance (Fig. 8-10). The further rise in the applied voltage leads to a nearly proportional increase in the device current.

The sudden rise in the resistance of the material in response to a strong electric field occurs at discrete points because the semiconductor material inevitably has discontinuities, and the transferred electrons form bunches known as *domains of space charge* (Fig. 8-11). The domains are usually formed near the cathode ("–") contact and migrate at a high velocity

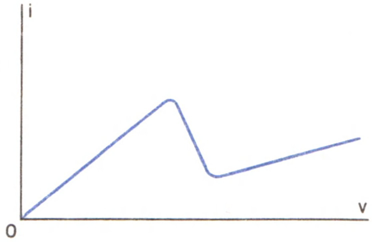


Fig. 8-10
Current-voltage characteristic of a Gunn diode

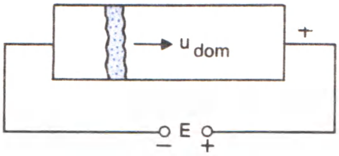


Fig. 8-11
Domain in a Gunn diode

towards the anode ("+") contact. Within each bunch, or domain, the electrons have a lower velocity than they do outside, and so the space charge density is increased there. The electric field in a domain is stronger than outside it where the electrons have a higher velocity. For this reason, the electrons to the right of a domain migrate faster towards the anode, leaving behind an electron-depleted region. On the left of a domain, however, new electrons arrive faster than others leave it. This process is responsible for the migration of domains from cathode to anode.

On reaching the anode, a domain disappears, and a new one emerges at the cathode, migrates towards the anode, and so on. This alternating growth and decay of domains is accompanied by periodic changes in the resistance of the Gunn diode which, in consequence, cause the diode current to oscillate. With a sufficiently small distance that has to be traversed by the domain (the anode-to-cathode spacing), the current oscillates at frequencies lying in the microwave region. This frequency is given by

$$f_{\text{Gunn}} = u_{\text{dom}}/L$$

where u_{dom} is the velocity at which the domain migrates (for gallium arsenide it is about 10^7 cm s^{-1}), and L is the device length (usually several micrometres).

It follows then that at $L = 10\ \mu\text{m}$,

$$f_{\text{Gunn}} = 10^7/10^{-3} = 10^{10}\ \text{Hz} = 10\ \text{GHz}$$

An important distinction of the Gunn diode is that all of the semiconductor material of the device, and not only a *p-n* junction (which accounts for a small proportion of the total bulk in other devices that use a *p-n* junction), contributes to device operation. Therefore, Gunn diodes are able to handle higher values of power. At present, Gunn diodes can generate CW oscillations at a power level of several tens of watts, and pulses at a power level of several kilowatts at an efficiency of several per cent to tens of per cent. Theoretically, it is quite feasible to build Gunn diodes for pulse powers of several hundred kilowatts at frequencies running into tens of gigahertz.

8-4 Heterojunction Devices

So far we have dealt with semiconductor devices using what are known as *homogeneous p-n junctions*, or *homojunctions*, that is, junctions formed between two similar semiconductor materials doped with opposite impurities (that is, donors and acceptors). There are also *heterogeneous junctions*, or *heterojunctions*, that is, junctions between two dissimilar semiconductors of opposite polarity types, differing in the bandgap width. A theory of such junctions was first advanced by A. I. Gubanov of the Soviet Union in 1951, and, somewhat later, by W. Schokley of the United States who took out a patent on the use of heterojunctions in semiconductor devices.

If we have two dissimilar semiconductors, labelled *1* and *2*, we can build four types of haterojunctions differing in the type of impurities (or doping) they contain, namely n_1-n_2, p_1-p_2, p_1-n_2, and p_2-n_1. The metal-semiconductor junction may be treated as a special case of the heterojunction. The most explored heterojunctions are Ge-GaAs, Ge-Si, GaAs-GaP, and GaAs-InAs.

Semiconductor devices using heterojunctions offer a number of advantages and are very promising in many respects. For one thing, diodes with n_1-n_2 or p_1-p_2 heterojunctions have a high operating speed and a high cutoff frequency because they are free from carrier storage – a relatively slow process so typical of conventional *p-n* junctions. Heterojunctions may require less than 1 ns in order to move from the ON state into the OFF state. Of special interest are high-power *laser-type heterodiodes* in which the generated power is emitted as a radiation rather than dissipated as heat in the device itself. *Tunnel heterodiodes* have a higher I_p/I_v ratio within the negative-resistance portion of their *V-I* characteristic (as compared with conventional tunnel diodes), are slow to age, and less susceptible to radiation damage than the usual tunnel diodes. The best performance at microwave frequencies has so far been shown by *avalanche transit-time heterodiodes.*

A good deal of interest has been shown in heterojunction transistors, such as p_1-n_2-p_2 types. They have a characteristically high value of alpha current gain, low emitter capacitance, and low series base resistance so that they are able to operate at higher cutoff frequencies. Some parameters can be improved by using a heterojunction as the gate in JFETs and in thyristors. Notably, thyristors have then a higher operating speed.

A number of important advantages are offered by heterojunction optoelectronic devices.

The crucial problem with regard to heterojunction devices at this writing is what are called as the interface-state effects, that is, the processes that occur at the boundary between the two semiconductors that make up a heterojunction. A careful selection of source materials and improvements in production methods are essential.

8-5 The Unijunction Transistor

The *unijunction transistor* (UJT), otherwise called the *double-base transistor*, is shown in Fig. 8-12. It has only one *p-n* junction and its structure resembles that of a JFET, but it operates on an entirely different principle. The *n*-region (the base), which has ohmic contacts B_1 and B_2 at its ends, is not a channel that would change its resistance due to changes in its cross-sectional area. The p^+-type emitter and the base form a p^+-n junction which is forward-biased externally, rather than reverse-biased as is a JFET. The output current flowing through the base produces across the portion from the emitter to contact B_1 a voltage drop V_{int} which acts as a reverse-bias voltage for the emitter and turns it off. When the external forward voltage equal to $E_1 + v_{in}$ exceeds V_{int}, the resultant voltage across the junction becomes forward-biasing, the junction is turned on, and holes are injected from emitter into base. As a result, the base resistance decreases. A change in the input voltage brings about a change in the injection level, in the base resistance and, as a consequence, in the output current so that an amplified voltage is developed across the load.

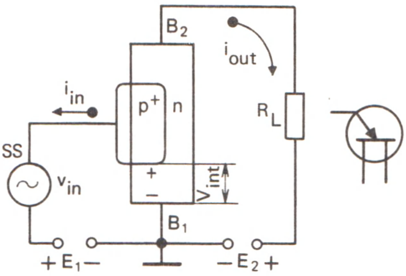


Fig. 8-12

Unijunction transistor (a double-base diode)

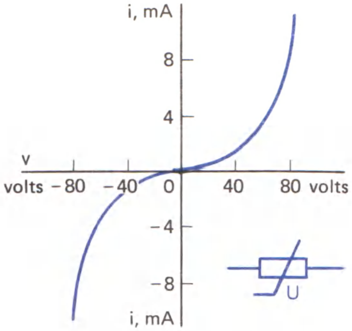


Fig. 8-13

Current-voltage characteristics and graphical symbol of a varistor

UJTs are most often used in timing, triggering, sensing, and waveform-generation circuits. In their frequency performance, however, they are inferior to conventional bipolar transistors and are able to operate at low frequencies only.

8-6 Semiconductor Resistors

Semiconductor resistors possessing nonlinear properties are called *varistors.* They are mainly fabricated from powdered silicon carbide, SiC, and some binder. Their nonlinear behaviour is mainly due to the heating of microcontacts between the SiC grains. Varistors may come as bars or discs. Their *V-I* characteristic and graphical symbol are shown in Fig. 8-13.

Varistors may be used on both d. c. and a. c. at frequencies up to several kilohertz. At higher frequencies, the inherent capacitance of the device begins to be felt. Practical uses for varistors are many and diverse: they are used for protection against overvoltages, in voltage regulators and limiters, and in various automatic-control circuits.

The key parameters of varistors are: the nonlinearity factor β defined as the d. c.-to-a. c. resistance ratio (usually it ranges between 3 and 6); the absolute maximum voltage rating (from tens of volts to several kilovolts); the absolute maximum power dissipation (1-3 W); the temperature coefficient of resistance ($-5 \times \times 10^{-3}$ K^{-1} on average); and the operating temperature limit (60-70°C).

Sometimes, semiconductor resistors are fabricated from materials that have a large, nonlinear negative temperature coefficient of resistance, in which case they are called *thermistors* (which is short for a temperature-sensitive resistor). Thermistors are usually shaped as rods, slabs, discs, or beads, and named accordingly. There are also thermally sensitive resistors which have a positive temperature coefficient of resistance; they are called *posistors*.

Figure 8-14 is a plot of the resistance of a thermistor as a function of temperature, a *V-I* characteristic for several conditions of heat withdrawal, and a circuit symbol for the device. Curve *1* holds for the worst case of heat abstraction, and curve *2*, for the best.

Thermally sensitive resistors are employed as temperature transducers and nonlinear resistors in various automatic control applications.

There are specially designed small-sized thermally sensitive resistors, which are used to measure radiant energy owing to their ability to absorb electromagnetic power. They are collectively called *bolometers*, of which the most predominant types are *thermistors* and *barretters*. Some thermistor bolometers have a heater so as to bring the device's temperature to a value where its sensitivity is at its highest.

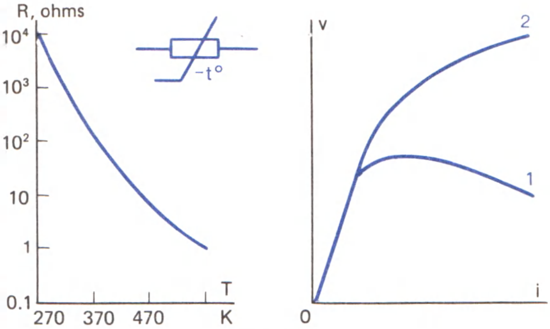


Fig. 8-14

Characteristics and graphical symbol of a thermistor

The key parameters of thermally sensitive resistors are:

- the rated resistance (from several ohms to several kilohms with tolerances of ± 5, ± 10, and ± 20%);
- the temperature coefficient of resistance, TCR, usually ranging from 0.8×10^{-2} to 6.0×10^{-2} K^{-1}.

Their data sheets may also specify the d. c. and a. c. resistance at some particular temperature (say, 20°C). For the device to show a consistent performance, the user should never exceed its maximum safe temperature and absolute maximum power dissipation rating.

Chapter Nine
An Outline of Integrated Circuits

9-1 General

Big strides in electronics and telecommunications have brought with them an extreme sophistication in the circuitry, more rigorous requirements for performance in general and reliability in particular, and the use of a huge number of circuit components. This is especially true of computers. Finally, a state has been reached where computers and telecommunications equipments could no longer be made solely of what are called *discrete components*–that is, components which have been fabricated prior to their installation (for example, diodes, transistors, resistors, capacitors, etc.). The situation can best be illustrated by the following example.

Suppose that we are to build a computer made up of 10^8 components, and each discrete part has, on the average, a mass of 1 g, a volume of 1 cm^3, a power dissipation of 10 mW, a failur e rate of 10^{-5} h^{-1}, and a cost of 50 kopeks. It follows then that a complete computer will weigh 100 metric tons, have a volume of 100 m^3, require 1 MW of power, show an entirely unacceptable failure rate of 10^3 h^{-1}, and will cost 50 million roubles. With a failure rate like that, this means that any one component of the computer would fail every 3 seconds. It is obvious that there is no sense in building such a computer, using discrete components. Such computers and telecommunications equipments comparable in complexity are currently built on the basis of *integrated circuits*, or ICs for short.

The transition to ICs has been a gradual one. At first, electronic equipments employing discrete components had come to use printed wiring boards instead of the older wiring technique by which discrete components were interconnected by hook-up wire. Typically, interconnections were printed on a laminated-plastic board as metal films in a desired wiring pattern, and the necessary discrete components were soldered at the designated points. This approach served to reduce the volume of equipment a good deal. Next came *modules* and *micromodules* – complete functional circuits (such as amplifiers, oscillators, converters, and the like) built into miniature cases or packages. Micromodules could be replaced in the case of failure with a minimum of delay. Miniature diodes, transistors, resistors, capacitors, inductors and other circuit components were designed specifically for such micromodules. Some micromodules even used miniature printed-wiring boards. Still, micromodules finally failed to solve the problem completely.

A real breakthrough in the design of extremely sophisticated electronic equipment came with the advent of integrated circuits. They are called 'integrated' because each complete circuit including its circuit elements and connections is manufactured as a single package.

The basic types of ICs are *film integrated circuits*, *monolithic integrated circuits*, *hybrid integrated circuits*, and *multichip integrated circuits*.

In a *film IC*, all the components and interconnections are fabricated in the form of various films (conducting, resistive, and dielectric) produced *in situ* on an insulating substrate. To define the nature of a film IC, additional modifiers may be added, such as a *thin-film* IC, a *thick-film* IC, etc.

In *a monolithic IC* all the circuit components are manufactured into or on top of a single chip of a semiconductor material (mostly, silicon). The individual parts of a monolithic IC are not separable from the complete circuit once formed. Interconnections between the various parts of a monolithic IC are made by a pattern of conducting material on the surface of the chip.

A *hybrid IC* is an arrangement which consists of several discrete components attached to a ceramic substrate and interconnected either by wire bonds or a suitable metallization pattern. The individual components are unencapsulated and may consist of diffused or thin-film components or one or more monolithic circuits. Once fabricated, an individual component cannot be altered without destroying the entire circuit.

The principal advantages of ICs are that they are small in size and weight, draw little power from supply sources, show a remarkable reliability owing to the reduced number of soldered joints, can operate at very high speeds because the interconnections between the components are short and the time it takes a signal to travel these distances is very small, and have

Table 9.1

NUMBER OF COMPONENTS AND RELIABILITY OF ELECTRONIC EQUIPMENT

Circuit type	Number of components per cm^3	Failure rate, h^{-1}
Tube circuits, discrete components	10^{-1}	10^{-5}
Transistor circuits, discrete components	1	10^{-6}
Micromodule circuits	10	10^{-7}
Film and hybrid integrated circuits	10^2	10^{-8}
Monolithic integrated circuits	10^{-3}-10^{-5}	10^{-9}

a relatively low cost. Table 9-1 compares several types of circuits in terms of packaging density (defined as the number of components or equivalent components in a unit volume of a circuit or package) and reliability. Of course, only average values are quoted. The human brain contains as many as 10^9 nerve cells (neurons) per cubic centimetre. As is seen, ICs still fall short of this figure by a wide margin. The reliability of the brain is very high, too. The brain is inferior, however, to ICs in terms of operating speed. The brain owes its high reliability to the fact that its cells duplicate one another (they provide what is called *redundancy* in reliability theory)—should one cell fail, its neighbour will take over at once.

Along with the unquestionable advantages, ICs suffer from some limitations. For one thing, they are low-power devices. High-power ICs are yet unfeasible. Problems also arise in making high-value integrated capacitances and inductances. Interconnections between the individual IC chips are made by the old techniques, take up a considerable space, and impair reliability.

The complexity of ICs that may be produced on a single chip of silicon is defined by descriptive terms. Thus we have *small-scale integration* (SSI) with no more than 10 components per chip; *medium-scale integration* (MSI) with 10 to 100 components per chip; *large-scale integration* (LSI) with 100 to 1000 components per chip; *very large scale integration* (VLSI) with over 1000 components per chip; and, finally, *extra large scale integration* (ELSI) with a still greater packaging density. LSI, VLSI and ELSI chips are often described by the number of bits that can be stored in a computer memory of similar complexity. Thus, LSI denotes circuits of complexity up to 16 kilobits; VLSI describes circuits with a capability between 16 kilobits and one megabit; and ELSI refers to circuits containing more than one megabit.

A further classification of ICs is into *digital* and *analog*. Digital ICs are mainly used in computers; they operate in the switching (pulse) mode and may, at any given instant, reside in one of two distinct states because present-day computers use the binary number system where only two digits are used, 0 and 1. Analog ICs operate under conditions in which currents and voltages vary continuously in a particular fashion, say sinusoidally. Analog ICs can operate as amplifiers, oscillators, various signal converters, etc.

9-2 Film and Hybrid Integrated Circuits

In a film IC, the substrate is made of an insulating material, such as glass, ceramics, etc. The passive elements, that is, resistors, capacitors, inductors, and interconnections take the form of various films deposited upon the substrate. The active elements (diodes and transistors) cannot be manufactured by film technology. Thus, film ICs contain solely passive elements.

As has been noted, there may be *thin-film* ICs in which the film thickness is not more than 1-2 µm, and *thick-film* ICs in which a thicker film is used. The difference between the two types is, however, not in film thickness, but in the technology used to apply the film.

The substrate of a film IC is an insulating board 0.5-1 mm thick, carefully ground and polished. *Film resistors* are manufactured by depositing a resistive film upon the substrate. If the resistance of a resistor is not to be very high, the film is a high-resistance alloy, such as Ni-Cr. High-value resistors are fabricated from *cermets* which are metal-dielectric mixtures (the first half of the term is derived from *cer*amic, and the second, from *met*al). The ends of the resistive film are fitted with contacts in the form of metal films which also serve as interconnections to other components. The resistance of a film resistor depends on the thickness and width of the film, its length and material. For higher values of resistances, film resistors can be made in a zig-zag pattern. The structure of film resistors is shown in Fig. 9-1.

The resistivity (called *sheet resistivity*) of film resistors is expressed in ohms per square (Ω/square) because the resistance of a given film shaped as a square is independent of the square

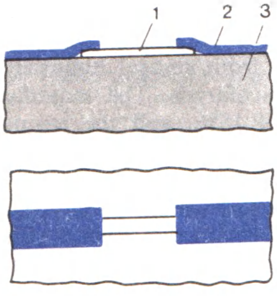


Fig. 9-1

Film resistor: (*1*) resistive film; (*2*) lead; (*3*) substrate size. To demonstrate, if we make the side of a square twice as great as it was before, the distance to be traversed by current will also be doubled, as will the cross-sectional area of the film; in consequence, its resistivity will remain unchanged.

Thin-film resistors are superior to thick-film devices in terms of accuracy and stability, but they are more difficult and expensive to make. Thin-film resistors may have a sheet resistivity of 10 to 300 Ω/square, and they may come with rated values of 10 to 10^6 Ω. The manufacturing tolerance is ± 5%; when properly trimmed, their values may be made accurate to within ± 0.05%. Trimming consists in that some of the resistive layer is removed in one way or another, and the resistance value deliberately made smaller than is actually needed is thus raised. The temperature coefficient of resistance for thin-film resistors is 0.25×10^{-4} K^{-1}. The value of these resistors changes very little with time in service.

Thick-film resistors have a sheet resistivity of 5 Ω/square to 1 MΩ/square, and come in values from 0.5 to 5×10^8 Ω. The untrimmed tolerance is ± 15%, and the trimmed tolerance is ± 0.2%. The TCR is about 2×10^{-4} K^{-1}. Their stability with time is inferior to that of thin-film resistors.

Film capacitors are most often manufactured with two plates. One plate is deposited on top of a substrate and is extended as an interconnection. It is then overlaid with a dielectric film which is topped by the second plate which likewise extends into an interconnection (Fig. 9-2). Similarly to film resistors, capacitors may be *thin-film* and *thick-film*. The dielectric used most often is the oxide of silicon, aluminium, or titanium. The capacitivity may range from tens to thousands of picofarads per square millimetre. Accordingly, a capacitor with a surface area of 25 mm^2 will have a rated capacitance from several hundred to tens of thousand picofarads. The manufacturing tolerance is ± 15%, and the temperature coefficient of capacitance is 0.05×10^{-4} to 0.2×10^{-4} K^{-1}.

Film inductors are manufactured as flat spirals, mostly rectangular in shape (Fig. 9-3). The width of conducting strips and the spacing between them is usually several tens of micrometres. This works out to an inductivity of 10-20 nH mm^{-2}. An area of 25 mm^2 yields an inductance of up to 0.5 μH. As a rule, film inductors are fabricated with a maximum inductance of several microhenrys. A greater inductance can be obtained by overlying the inductor with a ferromagnetic film to act as a core. There is a snag in making a terminal lead to the inner end of the film inductor. Therefore, a dielectric film is deposited at an appropriate spot of the inductor, and the dielectric is then overlaid with a metal film to serve as a terminal connection.

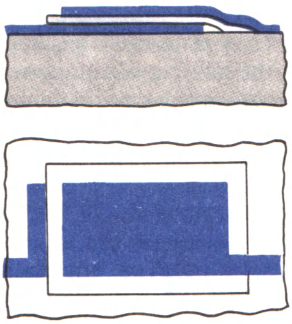


Fig. 9-2
Film capacitor

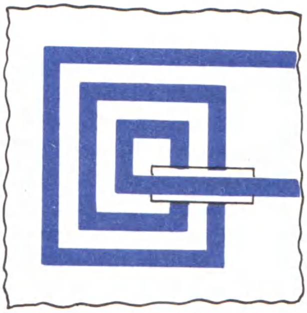


Fig. 9-3
Film inductor

Film ICs may be *RC*-networks such as *RC* filters or other circuits.

As already noted, in a *hybrid* IC the passive components are manufactured by film technology, and the active components (diodes and transistors) are discrete parts which are left unencapsulated and bonded to the substrate at the requisite points and connected by fine conductors to the film components of the IC. Sometimes, hybrid ICs may use some of their passive components in discrete form, such as miniature capacitors of a capacitance and miniature inductors of an inductance such that the

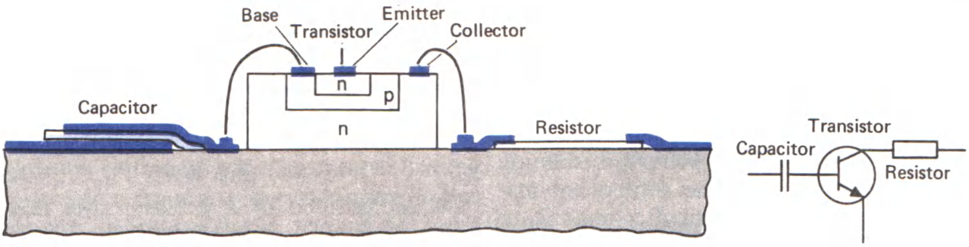


Fig. 9-4

Structure and diagram of a hybrid IC

respective parts cannot be manufactured by film technology. The discrete components of a hybrid IC may also include miniature transformers. In some cases, hybrid ICs may include complete monolithic ICs as discrete components of a much smaller size than the host hybrid IC itself.

A hybrid IC consisting of a capacitor, a transistor, and a resistor is shown in Fig. 9-4. This IC may be a part of an amplifying stage. The conductors made to the transistor or other discrete components are joined to the respective points in the circuit by any one of several connection techniques, the most popular one being *thermo-compression bonding.* A thermo-compression bonded connection can be produced between certain metals and semiconductors by the simultaneous application of heat and pressure.

A hybrid IC is fabricated in several steps. The first step is to make a substrate. It is thoroughly ground and polished. The next step is to deposit resistive films, the lower capacitor plates, inductors, and interconnections. Following that, dielectric (insulating) films are deposited, which are then topped by metal films again. Then comes the turn of discrete active and other devices, and their leads are joined to the respective points in the circuit. The IC is then packaged in a case and joined to the contact pins of the case. After appropriate tests, the case is hermetically sealed and the necessary symbols are marked on it.

9-3 Monolithic Integrated Circuits

As already explained, in a *monolithic* IC all the circuit components are manufactured into or on top of a single chip of silicon 200-300 μm thick. Chips for monolithic ICs usually measure from 1.5 × 1.5 to 6 × 6 mm. As compared with film and hybrid ICs, monolithic chips have the highest packaging density and the best reliability (the lowest failure rate). Among their limitations are the relatively low quality of passive elements (resistors and capacitors) and the fact that inductors cannot be made in a monolithic IC. Nevertheless, monolithic ICs are leading the field of microelectronics as the basis for LSI, VLSI and ELSI chips.

Isolation. Since all the circuit components are manufactured in a single chip of silicon, it is important to electrically separate or *isolate* them from one another. Two methods are used most often. The simplest and less expensive technique is to use a *p-n* junction – this known as *diode isolation.* For this purpose, *n*-type regions are produced on a *p*-type silicon chip or wafer by a process known as *isolation diffusion.* These *n*-regions are called *isolation islands,* or *isolated regions* (Fig. 9-5). They are then utilized to produce the desired passive or active components, and the *p-n* junctions between the isolation islands and the substrate in an operating IC are always reverse-biased. To this end a negative potential of several volts is always applied to the wafer. Under reverse bias, a silicon *p-n* junction presents a very high resistance (several megohms), and this resistance provides the desired isolation between circuit components. Obviously, the isolation resistance between any two circuit components will be equal to twice the reverse resistance of the isolating *p-n* junction. It should be kept in mind that each of these junctions has a barrier

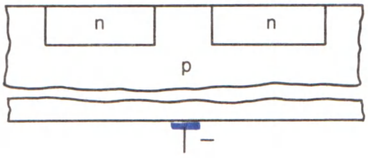


Fig. 9-5

Element isolation by a *p-n* junction (diode isolation) capacitance of its own, and so parasitic capacitive coupling exists between the circuit components owing to the junction capacitances.

The other method is by means of a dielectric layer, as shown in Fig. 9-6, and the technique is known as *dielectric isolation*. Here, too, isolation islands are produced for the subsequent manufacture of the required circuit components, but the islands are separated from the silicon chip by a thin layer of silicon dioxide, SiO_2. This SiO_2 layer complicates the manufacture of ICs, but it is more effective than an isolating *p-n* junction. Another advantage of dielectric isolation over diode isolation is the reduced parasitic capacitance between the islands and the wafer because the dielectric layer is several times as thick as the *p-n* junction. Because of complications in the manufacturing process, diode isolation is used more often than dielectric isolation. In our subsequent discussion, the diagrams will show the latter type of isolation, although there are several more isolation techniques apart from the two we have just described.

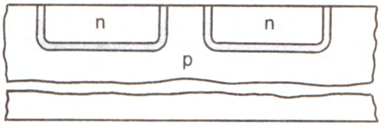


Fig. 9-6

Element isolation by a dielectric layer (dielectric isolation)

Bipolar transistors. They can be manufactured by planar or planar-epitaxial technology. The collector, base and emitter regions are formed in a silicon chip by diffusion as shown in Fig. 9-7. The diagram shows a sectional view and a plan view of the transistor. The transistor structure extends into the substrate to a depth of not more than 10-15 μm, and the linear dimensions of the transistor on the surface do not exceed several tens of micrometres.

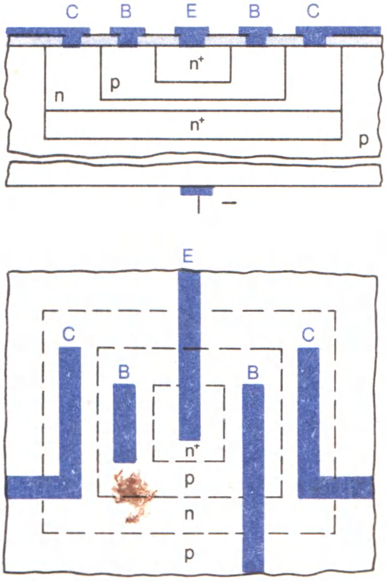


Fig. 9-7

Bipolar transistor of a monolithic IC

As a rule, *n-p-n* transistors are manufactured. What is known as a *buried layer* is made in the collector. It has a heavy impurity concentration (called n^+) and its objective is to reduce the resistance and, in consequence, the power lost in the collector region. At the collector junction, however, the collector region must have a reduced impurity concentration so that the junction could have an ample thickness. Then it will have a lower capacitance and a higher breakdown voltage. The emitter region is likewise the n^+-type very often in order to reduce its resistance and to step up carrier injection. A layer of silicon dioxide is formed over the entire surface, and two contacts are made to each of the collector and base so that the transistor can be connected to adjacent circuit components without cross-overs. The cross-overs are undesirable because they complicate the manufacture of ICs. Where one connection crosses another, it is necessary to deposit a dielectric film over the lower connection and to deposit the upper connection on top of the film – this involves two extra manufacturing steps. Also, cross-overs always present a risk of breakdown in the case of an inadvertent overvoltage.

An important problem in the manufacture of ICs is the component layout which must be such that no or a minimum number of cross-overs have to be made. Also, it is vitally important to minimize parasitic coupling between the circuit components as much as practicable. When an IC consists of a great number of circuit components, a huge number of likely layouts may exist, and a good deal of time has to be spent in order to pick an optimal layout. Of late, this job has been relegated to computers which are able to pick the best layout in the shortest time.

Typical bipolar transistors for monolithic ICs have a common-base (beta) current gain of 200, a cutoff frequency of up to 500 MHz, a collector capacitance of not over 0.5 pF, a collector breakdown voltage of 50 V, and an emitter breakdown voltage of 8 V. The sheet resistivity of the *p*- and *n*-regions is several hundred ohms per square, and that of the n^+-regions, not more than 15-20 Ω/square.

It is to be stressed that some unwanted circuit components may be formed in monolithic ICs. As is seen from Fig. 9-7, for example, there is, in addition to the *n-p-n* transistor formed in the *p*-type wafer, also a parasitic *p-n-p* transistor which is formed by both the wafer and the collector and base regions of the transistor. In turn, the *n-p-n* transistor and the wafer make up a parasitic *n-p-n-p* thyristor. Owing to the reverse voltage existing across the isolation junction, the parasitic transistor and thyristor are normally turned OFF. Should, however, noise pulses reach them, they may be turned ON and operate against the user's desire.

Multiemitter transistors. Apart from conventional transistors, digital ICs employ multiemitter transistors. Their structure and circuit symbol are shown in Fig. 9-8. The example shows a four-emitter transistor. It can be turned on by applying a forward voltage pulse to any of the four emitters. Each emitter is connected to a turn-on pulse source of its own. Importantly, a turn-off pulse from any one source may not break through to any other source because the emitter junctions inoperative at a given instant are reverse-biased.

For example, if a negative voltage pulse is fed to emitter E_1, the associated emitter junction is rendered conducting, and a collector current begins to flow in the transistor. The negative potential is transferred from E_1 to the *p*-type base and, since emitters E_2, E_3 and E_4 are at zero potential, they present a very high resistance, and the pulse applied to E_1 cannot enter the circuits of E_2, E_3 and E_4 because these circuits are isolated from one another. If the four turn-on pulse sources had been connected together to one emitter, there would have been no isolation. In such a case, isolation could be produced by placing a diode in the circuit of each input signal source, but this is far more complicated than to use a single multiemitter transistor.

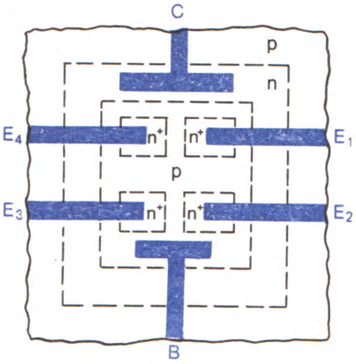
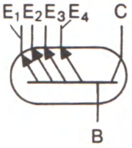


Fig. 9-8

Multi-emitter transistor



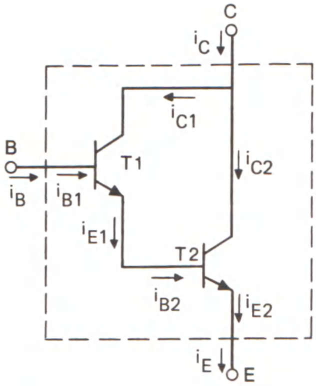


Fig. 9-9

Circuit of a compound transistor

It is to be noted that in a multiemitter transistor, the emitter in the ON state, the base, and an adjacent emitter make up, between them, a parasitic transistor. To minimize its effect, the adjacent emitters are spaced at least 10 μm apart so that a very wide base is produced in the parasitic transistor.

There are also *multicollector transistors* similar in the underlying principle to multiemitter transistors. Their structure can be visualized if, in the circuit of Fig. 9-8, we connect the four emitters as collectors, and operate the collector as an emitter.

The supergain transistor. Transistors thus named are sometimes used in integrated circuits. In them, the base is made a mere 0.2-0.3 μm wide so that the beta current gain may be increased by a factor of several thousand. However, the breakdown voltage of a supergain transistor is brought down to as low as 1.5-2 V.

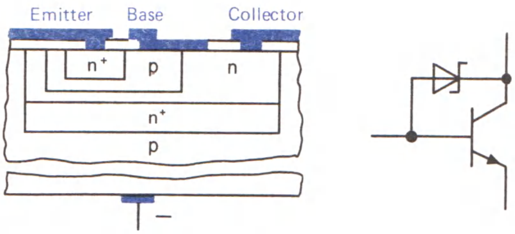


Fig. 9-10
Schottky barrier transistor

When the applied voltage is very high, a punch-through occurs – the collector-base depletion layer spreads through the entire base region and reaches the emitter junction so that a direct conducting path is formed from emitter to collector, and charge carriers from the emitter 'punch through' to the collector.

Compound transistors. These when used in ICs are each a pair of transistors so connected that a high-beta device results. Most often, use is made of what is known as the *Darlington pair* (Fig. 9-9). From inspection of the circuit diagram, we may write

$$i_{B2} = i_{E1} = (\beta_1 + 1)i_B \tag{9-1}$$

$$i_C = i_{C1} + i_{C2} = \beta_1 i_{B1} + \beta_2 i_{B2} \tag{9-2}$$

On substituting the expression for i_{B2} in Eq. (9-2) and dividing it by i_B, we obtain the resultant beta current gain of the compound transistor

$$\beta = \beta_1 + \beta_2 + \beta_1\beta_2 \approx \beta_1\beta_2 \tag{9-3}$$

When $\beta_1 = \beta_2 = 100$, we have $\beta = 10^4$. In practical compound transistors, β may be as high as several thousand.

The Schottky transistor. This is essentially a bipolar transistor in which a Schottky diode is used as a clamp between the base and collector. As has been shown in Sec. 2-4, the Schottky diode uses a metal-to-semiconductor contact and has rectifying properties. The device is advantageous in that it is free from diffusion capacitance and can therefore operate at frequencies as high as 3-15 GHz. Figure 9-10 shows the structure of the Schottky transistor in an IC and its circuit connection. It uses aluminium metallization to provide a nonrectifying, or ohmic, contact with the *p*-type base, but has a rectifying contact, that is, a Schottky diode, with the *n*-type collector. The Schottky transistor has a very high operating speed when used as a switch – it takes very little time for the device to change over from the ON to the OFF state.

Diodes (diode-connected transistors). Originally the diodes used in ICs were made as a structure having two regions of opposite conduction types, that is, similarly to the conventional *p-n* junction. Of late, the diodes utilized in ICs are made by using bipolar transistor structures in one of five possible connections. These five connections are shown in Fig. 9-11. The difference between them is mostly one of parameter values. In the BC-E connection, the base and collector are tied together. Here the reverse recovery time, that is, the time it takes the diode to move from the ON to the OFF state, is several nanoseconds. In the B-E connection, only the emitter junction is utilized, and the switching time is several times greater. The two connections have a minimal capacitance (a few tenths of a picofarad) and a minimal reverse current (0.5-1 nA), but their breakdown voltage is the lowest of all. However, the latter factor is of minor importance for low-voltage ICs. The BE-C connection, in which the emitter is tied to the base, and the B-C connection utilizing only the collector junction are about equal to the B-E connection in terms of the switching time and capacitance, but have a higher breakdown voltage (40-50 V) and a heavy reverse current (15-30 nA). The B-EC connection in which the two junctions are in parallel has the longest switching time (100 ns), the largest reverse current (as high as 40 nA), a somewhat greater

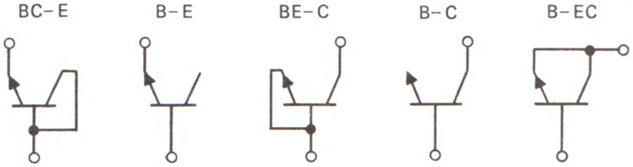


Fig. 9-11
Several schemes for using transistors as diodes

capacitance, and the same low breakdown voltage as the first two connections. The BC-E and B-E connections are used most often.

Some of the diode connections are used sometimes as voltage-reference (VR) diodes. Where a regulated (stabilized) supply of 5-10 V is required, resort can be made to the B-E connection reverse-biased into the breakdown region. Regulated voltages which are multiples of a forward voltage of 0.7 V can be supplied by a series combination of two BC-E connections (diodes) operating under forward bias. The temperature coefficient of voltage (TCV) of these VR diodes is several millivolts per kelvin. The lowest TCV is obtained with two n^+-p diodes connected in series opposition. In structure, they are similar to a transistor having two emitter regions (Fig. 9-12). One diode is reverse-biased into the breakdown region, and the other is forward-biased. Because the TCV under forward and reverse bias takes different signs, a compensating effect takes place so that the resultant TCV is less than 1 mV K^{-1}.

Integrated JFETs. Junction field-effect transistors can be manufactured on the same chip with bipolar transistors. Figure 9-13*a* shows the structure of a planar *n*-channel JFET. An *n*-type island has an n^+-region used as the drain and a *p*-region used as the gate. The drain is located at the centre, and the gate is arranged around it. Sometimes, a buried p^+-region is manufactured in order to minimize the initial channel width, but this complicates the production process. Figure 9-13*b* shows a *p*-channel JFET. Its structure is identical with that of a conventional *n-p-n* transistor. The job of the channel is done by the base.

Integrated MOSFETs. Metal-oxide semiconductor FETs are ousting bipolar transistors in ICs on an ever increasing scale. The reason is that MOSFETs offer a number of advantages over bipolar transistors. Among other things, they have a high input resistance and are simple in design. Enhancement-type MOSFETs are especially simple to manufacture. To make such a transistor, it suffices to diffuse an n^+-type source and an n^+-type drain into a *p*-type wafer (Fig. 9-14*a*). A reverse voltage is maintained across the junctions between these regions and the substrate, so that the transistors are isolated from the substrate and from one another. The channel is isolated from the substrate in a similar manner.

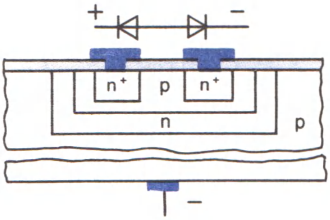


Fig. 9-12

Voltage-reference device made up of two diodes for temperature compensation

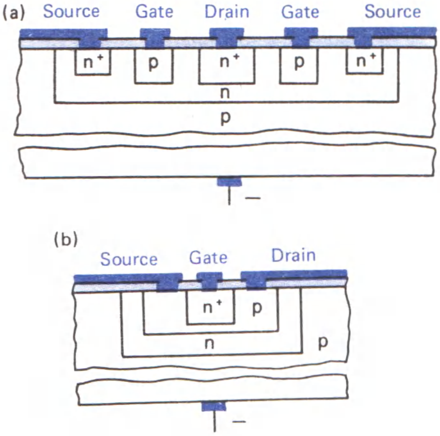


Fig. 9-13

Monolithic IC JFET: (*a*) *n*-channel and (*b*) *p*-channel

A more sophisticated method has to be used to make a *p*-channel MOSFET in a *p*-type substrate because an *n*-type isolation island must first be produced (Fig. 9-14*b*). Some ICs use pairs of *n*- and *p*-channel MOSFETs, usually referred to as complementary MOS field-effect transistors (COSMOS or CMOS for short, the latter abbreviation being most common). CMOSs are used in switching (digital) circuits and draw very little power. Some ICs use bipolar and MOS field-effect transistors manufactured on the same chip.

The fabrication of MOS transistors is being improved all the time, and quite a number of their structures have been developed at this writing. The principle of operation, however, remains the same in all of them.

Integrated resistors. A resistor in a monolithic IC is very often obtained by utilizing the bulk resistivity of diffused areas. The *p*-type base diffusion is most commonly used, although the *n*-type emitter diffusion is also employed. Figure

9-15 shows the structures of such integrated resistors. The resistance of a diffused resistor depends on the length, width, and thickness of the region doing the job of a resistor, and the bulk resistivity or, which is the same, the impurity concentration (or the doping level). A *p*-type integrated resistor (Fig. 9.15*a*) is made at the same time with transistor bases. The resultant sheet resistivity is then hundreds of ohms per square, and resistors can be obtained with values as high as tens of kilohms. Sometimes, resistors are fabricated in a zigzag pattern so as to increase their resistance. Where relatively low values of resistance are needed (units or tens of ohms), resistors are fabricated at the same time with the *n*-type emitters of transistors (Fig. 9-15*b*). The temperature coefficient of diffused integrated resistors is a few tenths of one per cent per kelvin and even less. The tolerance is ± (15-20)% or even greater.

A more recent trend has been to replace the diffusion process with *ion implantation*. It consists in that ions are implanted into the lattice of a semiconductor crystal by bombarding its surface with impurity ions under controlled conditions so that they penetrate into the crystal to a depth of 0.2-0.3 μm. The sheet resistivity of such resistors can be as high as 10-20 kilohms per square, and the resultant resistors may have resistance values as high as several hundred kilohms with a tolerance of not over ± (5-10)%.

It is important to note that integrated resistors have a parasitic capacitance to the substrate. Also, a resistor and its substrate form between them a parasitic *p-n-p* transistor. When designing an IC, the operating conditions for the transistors should be chosen such that the parasitic transistor will be turned OFF and have practically no detrimental effect. The parasitic capacitance to the chip limits the operating frequencies at which an integrated resistor may be regarded as being purely resistive. At frequencies exceeding some critical value, an integrated resistor may present an impedance which also includes a reactive component.

The job of an integrated resistor can be done by the channel of a MOS structure. Such MOS resistors are fabricated simultaneously with MOS transistors. If their structure is the same as that of the transistors, the desired resistance can be set by adjusting the gate voltage.

The pinch resistor. In structure, it is similar to a JFET. The job of a resistor is done by the channel, and the desired resistance is adjusted by varying the gate voltage – thus the pinch resistor operates much as the MOS-resistor does.

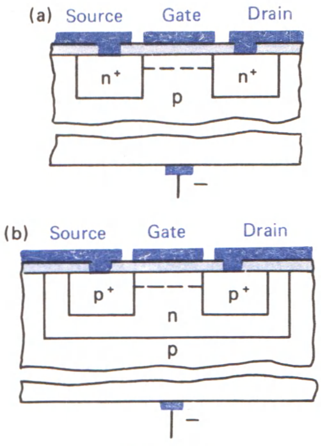


Fig. 9-14
Enhancement-type MOS-transistor of a monolithic IC: (*a*) *n*-channel and (*b*) *p*-channel

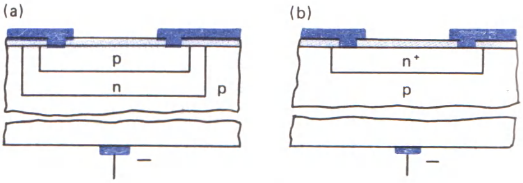


Fig. 9-15
Diffused integrated resistors in monolithic ICs

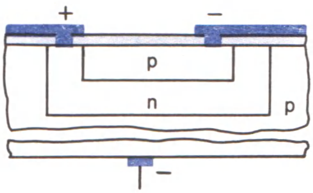


Fig. 9-16
Diffused integrated MOS-capacitor

Integrated capacitors. Typically, ICs use *diffused capacitors* which utilize the barrier capacitance of a *p-n* junction. The capacitance of such a *junction capacitor* whose structure is shown in Fig. 9-16 depends on the junction area, the permittivity of the semiconductor, and the junction width (or thickness) which is in turn a function of the impurity concentration (the doping level). Where a high value of capacitance is essential, a junction capacitor is manufactured at the same time with the emitter junctions of transistors (that is, by the n-type emitter diffusion). Since the emitter region has the n^+-type of conduction, the capacitor junction will be thin; its capacitivity is relatively high, being about 1000 pF mm^{-2}. Junction capacitors may have rated values up to 1500 pF with a tolerance of $\pm$ 20%. The TCC is about -10^{-3} K^{-1}, and the breakdown voltage does not exceed 10 V. Unfortunately, these capacitors have a low figure of merit – of the order of not over 20 at 1 MHz. Junction capacitors manufactured during the collector dffusion have a lower capacitivity (about 150 pF mm^{-2}) and will usually have a capacitance of not over 500 pF with a tolerance of $\pm$ 20%. Their breakdown voltage is up to 50 V, their TCC is -10^{-3} K^{-1}, and the Q-factor at 1 MHz is 50-100. The relatively low Q-factor of the above junction capacitors is explained by the fact that their dielectric is a semiconductor p-n junction in which a good deal of power is lost. The plates, which are semiconductor layers, present a noticeable resistance as well.

Diffused integrated capacitors can only operate when reverse-biased, and the bias voltage must be held constant so as to assure a constant value of capacitance. Because the junction capacitance is a nonlinear quantity, diffused integrated capacitors can be used as variable capacitors controlled by varying the d. c. voltage across the device. When the reverse voltage is varied anywhere between 1 and 10 V, the capacitance will be varied by a factor of 2-2.5. Some electronic circuits use nonlinear capacitors, and their job can well be done by diffused integrated capacitors.

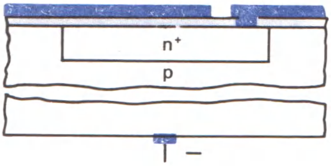


Fig. 9-17

MOS capacitor used in monolithic IC

Figure 9-17 shows a *MOS capacitor* used in monolithic ICs, especially those employing MOSFETs. One plate is a diffused layer of n^+-type silicon on which a thin layer of SiO_2 is produced. The SiO_2 layer is topped by a metal (aluminium) film acting as the other plate. The capacitivity is 300-400 pF mm^{-2}, and the capacitance values can be up to 500 pF with a tolerance of $\pm$ 25%. The breakdown voltage may be as high as 20 V. Among the advantages of MOS capacitors are a relatively low TCC (about 2×10^{-4} K^{-1}), a higher Q-factor (up to 200-250), and the ability to operate on a voltage of any polarity. The capacitance of MOS capacitors is less dependent on voltage than in diffused capacitors. As with other integrated circuit components, parasitic capacitances with respect to the substrate and parasitic transistors are formed in ICs using integrated capacitors.

Figure 9-18 shows a partial sectional view of a monolithic IC corresponding to the circuit diagram of Fig. 9-4, that is, consisting of a diffused capacitor, an integrated transistor, and an integrated resistor.

Integrated inductors. There is no way of producing an integrated inductor in a monolithic IC. Therefore, it is usual to design an IC so that it will need no inductors. Still, cases may arise where one needs an inductor or inductors. The need is satisfied by using an equivalent inductance which consists of an integrated tran-

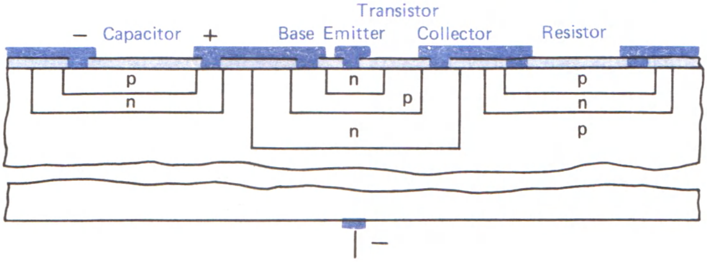


Fig. 9-18

Monolithic IC

sistor, an integrated resistor, and an integrated capacitor. An example of such an arrangement is shown in Fig. 9-19. Here an a.c. voltage V is applied between the collector and emitter of the transistor. For simplicity, the circuit feeding the d.c. supply voltage to the transistor is not shown. Some of the applied a.c. voltage V is fed via an RC-network to the base. The values of R and C are chosen so that $R \gg 1/\omega C$. Then I_{RC}, or the current in the RC-network may approximately be taken to be in phase with V. However, the voltage across the capacitor, V_C, lags behind I_{RC} by 90° in phase (it is said to be in phase-quadrature lagging). V_C is applied to the base and controls the transistor's collector current I_C which is in phase with V_C but is in phase quadrature lagging with V.

Thus, the transistor in the above circuit presents to V an impedance equivalent to the inductive reactance

$$x_L = V/I_C = \omega L_{eq}$$

In other words, the transistor is equivalent to some inductance

$$L_{eq} = V/\omega I_C$$

By varying the supply voltages so as to obtain the desired I_C, it is possible to obtain any desired value of L_{eq}. Since the impedance of the RC-network is many times x_L, its effect may be neglected.

Redundance. As has been noted, ICs have a very high reliability. Still, it must be enhanced even more in some, especially critical applications. One approach is through *redundance by duplication*. An example will illustrate how a diode can be duplicated. Figure 9-20 shows the connection of four diodes instead of one, with the diodes connected in both series and parallel. A diode may usually fail due to a breakdown (a short-circuit) or an open-circuit. Let each diode have a forward resistance of 10 Ω and a reverse resistance of *1* MΩ. If, say, diode *1* breaks down so that its resistance falls practically to zero, diodes *2* and *4* will operate so that their total forward resistance is 5 Ω and their total reverse resistance is 0.5 MΩ. If, on the other hand, diode *1* is open-circuited, diodes *2*, *3* and *4* will remain operating. They will present a total forward resistance of 15 Ω and a total reverse resistance of 1.5 MΩ.

A similar situation will arise, should any other diode fail. The probability that two diodes will fail at the same time and lead to a failure of all the four diodes is negligibly small. Thus, duplication is a very efficient way of improving reliability. (The same principle operates, as has been noted earlier, in the human brain.)

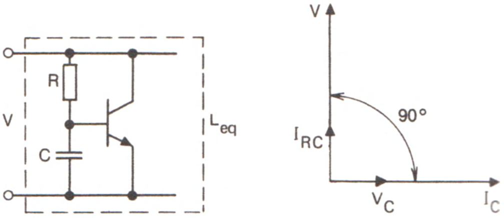


Fig. 9-19
Equivalent inductance

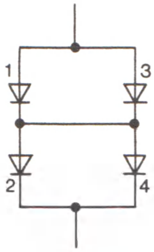


Fig. 9-20
Redundancy by duplication in the case of crystal diodes

Monolithic/thin-film ICs. In applications calling for better integrated passive elements, it is usual to combine monolithic and thin-film integrated circuits. This combination uses a semiconductor chip in which active elements (diodes and transistors) are manufactured. Then the chip is overlaid with an insulating SiO_2 layer on which thin-film passive elements are deposited. The circuit of Fig. 9-4 manufactured by this technology is shown in Fig. 9-21. Of course, such ICs are more expensive and larger in size than monolithic ICs, but they show a better performance.

Steps involved in the manufacture of ICs. In simplified form, the steps involved in the manufacture of monolithic ICs are shown in Fig. 9-22. Several more steps actually precede the sequence shown. Namely, a p-type single crystal ingot of silicon is first grown, the ingot is sliced into round wafers to form the substrate upon which all integrated components will be fabricated, one side of each wafer is lapped and polished to eliminate surface imperfections be-

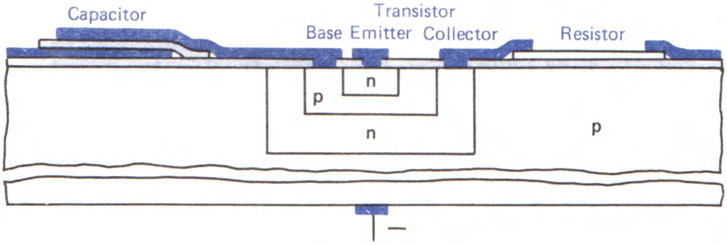


Fig. 9-21
Monolithic-thin film IC

fore proceeding with the next process, an *n*-type epitaxial layer is grown into the *p*-type substrate, this layer is polished and cleaned and a thin layer of silicon dioxide (*2*) is formed by exposing the epitaxial layer to an oxygen and steam atmosphere at about 1000°C. The monolithic technique requires the selective removal of the SiO_2 to form openings through which impurities may be diffused. For this purpose, the wafer is coated with a uniform film of a photosensitive emulsion called the *photoresist* (*1*).

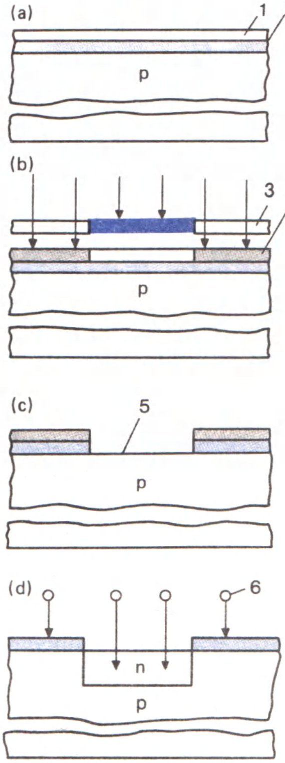


Fig. 9-22
Manufacture of a monolithic IC

The photoresist, when exposed to radiation, may become either acid-resistant (or, conversely, acid-soluble, depending on the choice of the material). Then the photoresist is irradiated with ultraviolet light (Fig. 9-22*b*) through a mask (*3*) prepared from a large black-and-white layout of the desired pattern of openings to give a negative or stencil consisting of transparent and opaque areas. The irradiated areas (*4*) of the photoresist become acid-resistant. Then the areas not exposed to UV radiation (*1* and *2*) are etched away (Fig. 9-22*c*) to leave openings (*5*) through which the dopants (*6*) are to be diffused from a hot gas (Fig. 9-22*d*). This is known as the *photolithographic process*.

The diffusion of the dopants through the openings produces *n*-type isolation islands. The photoresist remaining on the wafer is then removed (stripped) with a chemical solvent. Then everything is repeated all over again, that is, a SiO_2 layer is formed, a photoresist is applied, another photomask is laid over it, the photoresist is exposed to UV light to form a smaller opening through which acceptor atoms are diffused so as to form a *p*-region within the *n*-region, etc.

9-4 Charge-Coupled Devices and Integrated Injection Logic (I^2L) Chips

Charge-coupled devices (CCDs) belong to a wider class of what are called *charge-transfer devices* (CTDs). To visualize their operation, imagine a MOS transistor with an extremely long channel and with many gates closely spaced between source and drain. Such an arrangement is in effect a chain of MOS capacitors each of which is formed by one of the gates and the substrate, and the input signal is transferred along this chain.

In recent years, CCDs have come to be used in microelectronics as memory elements, delay lines, filters, signal-processing circuits, logic gates, and solid-state imaging devices (SSIDs) used instead of conventional TV tubes.

The first CCD was fabricated by Boyle and Smith of the United States in 1969. The idea of a memory made up of a chain of capacitors was put forward as far back as 1934 by V. K. Zworykin of the United States known for his work on TV camera tubes. In 1948, N. Wiener, the founder of cybernetics, also pointed to the possibility of storing information as a charge on capacitors and of transferring this charge from one capacitor to the next.

A major advantage of CCDs is their simple design. The general arrangement of a CCD is shown in Fig. 9-23. This is a three-phase symmetrical CCD consisting of a chain of MOS capacitors on a common *p*-type substrate. At the input and output, this chain may have diodes or FETs. The metal electrodes positioned along the device measure 10 to 15 μm and the spacing between them is 2 to 4 μm. The row of metal electrodes is separated from the *p*-type substrate by a layer of SiO_2 with a thickness of 0.1 μm.

A CCD can operate in two basic modes: (1) storage of information as a charge on one or several capacitors and (2) transfer of this charge from one capacitor to the next along the chain. In digital circuits, information is stored in binary form, that is, the presence of a charge corresponds to a logical 1, and the absence of a charge to a logical 0. In analog circuits, the amount of charge varies in a manner replicating the input signal.

Electrodes *1*, *2* and *3* make up a single cell of a CCD. The input electrode injects electrons across the n^+-p junction. The injection current can be controlled by varying the gate voltage. Charge transfer from electrode *1* to electrode *2* and so on can take place only if the voltage applied to the electrodes is positive and the voltage at the next electrode exceeds that at the preceding one or is equal to it, that is, if $V_2 \geqslant V_1 > 0$, then $V_3 \geqslant V_2 > 0$, and so on. The voltage pulses applied to the electrodes for charge transfer are called *clock pulses*. They are usually 10-20 V in magnitude.

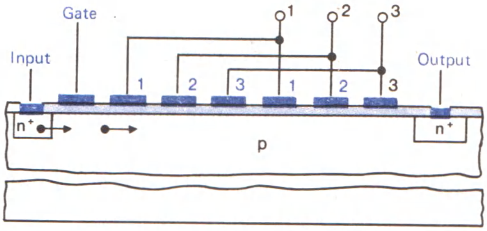


Fig. 9-23
IC using coupled-charge devices (CCDs)

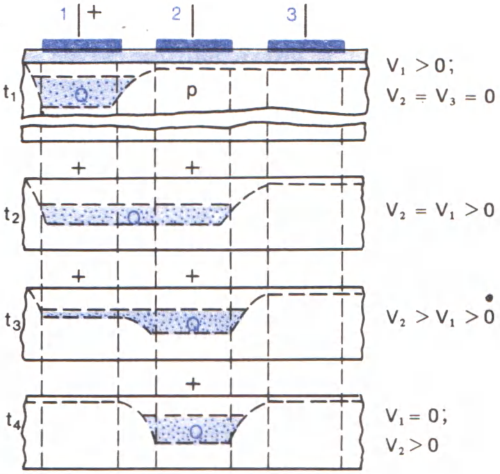


Fig. 9-24
Charge transfer in CCDs

Charge transfer from one electrode to the next is illustrated in Fig. 9-24. At time t_1, the charge packet Q is stored under electrode *1*, so $V_1 > 0$ and $V_2 = V_3 = 0$. The next instant, that is, at t_2, a voltage V_2 equal to V_1 is applied to electrode *2*. Now some of the charge packet is transferred into the region under electrode *2*. At t_3, when a voltage exceeding V_1 is fed to electrode *2*, that is, when $V_2 > V_1$, more of the charge packet moves into the region under electrode *2*. Finally, at instant t_4, when V_1 falls to zero and V_2 remains positive, $V_2 > 0$, all of the charge packet moves into the region under electrode *2*. Charge transfer from electrode *2* to electrode *3* and so on proceeds in a similar manner.

Charge storage in a CCD is limited by thermal electron-hole pair generation. The carriers thus produced are trapped in empty potential-energy wells and, in time, change the logic state from a 0 to a 1. This phenomenon, called the *dark-current effect*, sets the lower frequency limit for CCDs (it may be tens of kilohertz) and the maximum storage time (which does not exceed 100 μs). The dark-current effect can be minimized by cooling the device.

The operating speed of a CCD is limited by the fact that, although charge transfer from one electrode to the next takes place by the drift mechanism, the process terminates as a diffusion of carriers, and the diffusion velocity is substantially smaller than the drift velocity. Charge transfer is accompanied by a loss, but the efficiency is nevertheless as high as 97.0% to 99.9%. As the efficiency approaches unity (100%), the operating speed falls off. Practically, the operating speed of CCDs corresponds to a maximum frequency of 1 GHz.

For the simple planar electrodes of Fig. 9-23 or 9-24 it is necessary to use three-phase clocks to transfer the charge longitudinally in one direction only. However, two-phase clocking is possible if nonplanar electrodes are used and each CCD cell has two instead of three electrodes.

Of special interest are CCD-based solid-state imaging arrays. These imagers are simple in design and to make, small in size, light in weight, dissipate little power, have a high sensitivity, and are able to operate in visible, infrared and ultraviolet light. Despite some difficulties in manufacture, such imagers are very promising. Those already in use present an image as an array of several hundred thousand pixels ('pixel' is short for 'picture element'). Importantly, a CCD camera is the size of a cigarette pack, weighs as little as 200 g, and draws a mere 2 W of power.

Integrated injection logic (I^2L) was proposed almost simultaneously in 1971 by engineers at the IBM Laboratories (at Boeblingen, West Germany) and at the Philips Research Laboratories (at Eindhoven, Holland). The former group called their invention *merged-transistor logic* (MTL), and the latter designated it *integrated injection logic*. These two approaches led to essentially the same LSI device whose most popular acronym is I^2L. They offer a number of advantages over conventional ICs. For one thing, they permit a very high scale of integration. For another, the supply voltage may be as low as 0.5 to 0.9 V. A mere 10^{-12} J of energy is enough to do the necessary switching. I^2L chips are very simple because they consist of only bipolar transistors and need no other circuit components or component isolation. I^2L units can operate on currents and voltages varying between broad limits without causing any disturbance in the operation of the circuit.

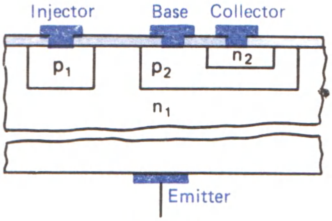


Fig. 9-25
Planar I^2L transistor

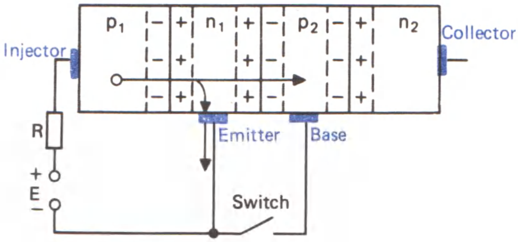


Fig. 9-26
Connection of an I^2L transistor in a circuit

Their temperature range extends from − 60° to + 125°C which is a very wide range indeed. They are simpler to make than any other devices based on bipolar transistors. They are especially suited for the manufacture of LSI chips, including those for microprocessors.

The principle on which an I^2L unit is built and operates is illustrated in Figs. 9-25 and 9-26. Figure 9-25 shows the structure of an n_1-p_2-n_2 planar I^2L transistor which is a unit cell of an I^2L chip. The n_1-type material in this transistor acts as an emitter. In contrast to conventional planar transistor, the device in our case is used in the inverse mode. In addition to the transistor, there is also a p_1-region called the *injector*, *Inj*. Together with the n_1-type emitter and the p_2-type base of the n_1-p_2-n_2 transistor, the injector makes up a p_1-n_1-p_2 transistor where the injector operates as an emitter. The junction between the injector and the substrate is called the *injector junction*.

It is convenient to trace the operation of an I^2L transistor by reference to Fig. 9-26 where the "+" and "−" signs at each junction label the space charges produced by impurity atoms. Let a forward voltage be applied to the injector junction from a supply source, *E*. The resistor *R* is included to limit the voltage across and the current through the injector junction. The applied forward voltage causes holes to move from the injector into the n_1-type emitter region, while the excess charge due to holes injected there under the influence of E is neutralized by in-coming electrons. The excess electrons and holes diffuse through the emitter to the n_1-p_2 emitter junction which is enriched with charge carriers so that its potential barrier is brought down, and its resistance decreases. Holes and electrons diffuse farther through the p_2-type base to the p_2-n_2 collector junction and likewise reduce its potential barrier and resistance.

In this way, the resistance of both junctions in the n_1-p_2-n_2 transistor is reduced, and the transistor operates close to saturation, that is, similarly to a closed switch. If we now short-circuit the base to the emitter by closing the switch *Sw*, the voltage across the emitter junction will fall to zero, no more carriers will be able to arrive at the p_2-n_2 collector junction, and its resistance will go up abruptly. Thus, the transistor will move to a state close to cutoff, and this corresponds to an open switch. The job of switch *Sw* may be done by another I^2L transistor driven to saturation.

A typical I^2L chip contains a chain of several transistors like those we have just examined. To all of them, the *n*-type chip is a common emitter, and there is a *p*-type region acting as a common injector. This explains the term 'merged'. By definition, if one semiconductor region is part of two or more devices, these components are said to be merged. In the chain making an I^2L chip, transistors at saturation alternate with transistors at cutoff. If any one of them moves to saturation, it short-circuits the base and emitter of the next transistor which is then driven to cutoff.

At present, quite a number of I^2L types have been developed, and there is a continuous advance in this direction of microelectronics.

Part Two

Electron Tubes

Chapter Ten

Behaviour of Electrons in Electric and Magnetic Fields

10-1 The Motion of Electrons in a Uniform Electric Field

The interaction of moving electrons with an electric field is the basic process in electron tubes. Therefore, before we go any further, it is important to examine the motion of an electron in a uniform and time-invariant electric field.

The laws governing the motion of a single electron in a uniform electric field may to a certain approximation be extended to the motion of electrons as an electron beam, if we neglect their mutual repulsion. This assumption does not lead to an appreciable error.

In most cases, the electric field is other than uniform; rather, it presents a fairly complicated pattern. The motion of electrons in a nonuniform electric field is difficult to investigate and belongs to a division of electronics called electron optics. If an electric field is only slightly nonuniform, we may well deem that electrons move in accord with the laws deduced for a uniform field. Using these laws, we may also consider, albeit in approximate terms, the motion of electrons in markedly nonuniform fields.

As will be recalled, an electron is a material particle that has a negative charge, e, of 1.6×10^{-19} coulomb, and a rest mass of $m = 9.1 \times 10^{-28}$ g. The mass of an electron increases with increasing velocity. Theoretically, at a velocity equal to that of light, or $c = 3 \times 10^{8}$ m s^{-1}, the mass of an electron must become infinitely large. In ordinary electron tubes, electrons move at a velocity of not more than $0.1c$, so the mass of electrons may be taken as constant.

The motion of an electron in an accelerating field. The sketch in Fig. 10-1 shows lines of force representing a uniform electric field between two electrodes which may be, say, the cathode and anode of a diode.

If the potential difference between the electrodes is $V$and the spacing between them is d, the field strength will be given by

$$E = V/d \qquad (10\text{-}1)$$

For a uniform field, E is a constant quantity.

Let the electrode at a lower potential, such as the cathode K, emit an electron whose kinetic energy is W_0 and whose initial velocity is u_0, directed along the lines of force, or flux. The electron is attracted by the electrode having a higher potential, that is, the anode A. Thus, the field accelerates the electron in its travel to the anode, and the field is quite aptly called an *accelerating* one.

The field strength (or intensity) is numerically equal to the force exerted on a unit positive charge. Therefore, the force acting on an electron is

$$F = -eE \qquad (10\text{-}2)$$

The "$-$" sign implies that the force F points in the direction opposite to the vector $\bar{E}$. Sometimes, the "$-$" sign is omitted.

Owing to the constant force F, an electron is given an acceleration

$$a = F/m$$

Moving rectilinearly, an electron acquires a maximum speed u and a maximum kinetic energy Wat the end of its travel, that is, when it strikes the collecting electrode. Thus, in an accelerating field, the kinetic energy of an electron is enhanced owing to the work done by the field in moving the electron. By the law of conservation of energy, the increase in the kinetic energy of an electron, $W - W_0$, is equal to the work done by the field, which is in turn defined as the product of the moved charge e by the potential difference through which it has been moved, V:

$$W - W_0 = mu^2/2 - mu_0^2/2 = eV \qquad (10\text{-}3)$$

If the initial velocity of an electron is zero,

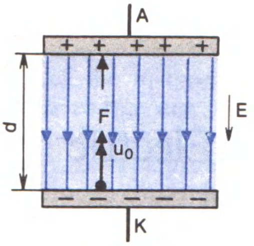


Fig. 10-1

Motion of electrons in an accelerating electric field

then

$$W_0 = mu_0^2/2 = 0 \text{ and } W = mu^2/2 = eV \quad (10\text{-}4)$$

that is, the kinetic energy of an electron is equal to the work done by the field.

With some approximation, Eq. (10-4) may be used even when the initial velocity u_0 is a small fraction of the final or terminal velocity u because

$$mu_0^2/2 \ll mu^2/2$$

If we agree that the charge on an electron is the unit of the quantity of electricity, then the energy acquired by an electron when passing freely through a potential difference of one volt, $V = 1$ volt, may be taken as the unit of energy called the *electron-volt*, eV. It is convenient to express electron energy in electron-volts and not in joules.

As follows from Eq. (10-4), the terminal velocity of an electron is

$$u = (2eV/m)^{1/2} \quad (10\text{-}5)$$

On substituting the expressions for e and m in the above equation, we can derive an expression for velocity, convenient for use in calculations, in metres or kilometres per second:

$$u \approx 6 \times 10^5 V^{1/2} \quad \text{or} \quad u \approx 600\ V^{1/2} \quad (10\text{-}6)$$

Thus, the velocity of an electron in an accelerating field is a function of the potential difference through which the electron has passed (or fallen, as is customary to say).

It is convenient to express the energy acquired by an electron in electron-volts subject to the equality

$$W_0 = eV_0 \quad (10\text{-}7)$$

that is, assuming that this energy is due to an accelerating field with a potential difference V_0.

Electrons can acquire appreciable velocities even in the presence of a small accelerating potential difference. For example, when $V = 1$ volt, the velocity of electrons is 600 km s^{-1}, while at $V = 100$ volts, the velocity is 6 000 km s^{-1}.

The time it takes for an electron to cover the spacing between the electrodes is given by

$$t = d/u_{av} \quad (10\text{-}8)$$

where u_{av} is the average velocity or, rather, arithmetic mean of the initial and terminal velocities

$$u_{av} = (u_0 + u)/2 \quad (10\text{-}9)$$

If $u_0 \ll u$, then

$$u_{av} \approx u/2 \text{ and } t \approx 2d/u \quad (10\text{-}10)$$

On substituting here the expression for the terminal velocity, we obtain the transit time in seconds

$$t = 2d/(6 \times 10^5 \times V^{1/2})$$
$$= 0.33 \times 10^{-5} d/V^{1/2} \quad (10\text{-}11)$$

where the spacing d is in metres. If we express it in millimetres, we finally get

$$t = 0.33 \times 10^{-8} d/V^{1/2} \quad (10\text{-}12)$$

For example, at $d = 3$ mm and $V = 100$ volts, the transit time is

$$t = 0.33 \times 10^{-8} \times 3/100^{1/2} = 10^{-9} \text{ s}$$
$$= 10^{-3}\ \mu\text{s} = 1 \text{ ns}$$

The transit time of electrons in electron tubes is more difficult to find because the field in a tube is anything but uniform. Practically, the transit time is 10^{-8}-10^{-10} s. This is a very short time and it may be neglected in many cases so that electron tubes may be deemed free from transit-time (electron inertia) effects. Still, since electrons have a finite mass, they cannot change their velocity instantaneously and cannot move from one electrode to the other likewise instantaneously. At UHF and SHF, the transit time of electrons in a tube becomes comparable with the period of oscillation. For example, at $f = 1000$ MHz, $T = 10^{-9}$ s, and an electron tube ceases to be free from transit-time (electron inertia) effects. Thus the electron inertia becomes a serious factor. At LF and HF, the period T is many times the transit time of an electron and the a.c. voltages at the tube electrodes remain practically unchanged in the meantime. Therefore, it may be thought that during the electron travel from one electrode to the other the electrode voltages remain constant.

The operation of a tube at constant electrode voltages is referred to as the *static mode of operation*. When the voltage of at least one electrode varies so fast that the relationships of the static mode are no longer applicable, a tube is said to be in the *dynamic mode of operation*. If, on the other hand, the voltage of at least one electrode varies at a low frequency so that the events occurring in the tube may be treated, at least approximately, on the basis of static relations, we have what is often called the *quasi-static mode of operation*.

The expressions describing the energy, velocity and transit time of an electron hold for any part of its path, but the quantities W, u, t, d and V refer to a particular part of the path rather than to the entire interelectrode spacing. While the field strength varies from one part of the path to another, the terminal velocity of an electron is solely decided by the terminal potential difference and the initial velocity of the electron. Although its acceleration will be different at different points along the path, this factor is of no significance for the calculation of the terminal velocity. It follows from the law of conservation of energy that the terminal potential difference V is equal to the algebraic sum of the potential differences existing at various points along the path. Therefore, the overall increment in the kinetic energy is equal to the product eV.

The difference in electron acceleration between the various points along the path tells only on the transit time. For the entire path it is equal to the sum of the time intervals during which an electron traverses the individual segments of the path.

The motion of an electron in a retarding field. Let the initial velocity u_0 of an electron point in a direction opposite to that of the force F exerted on the electron by the field (Fig. 10-2). In other words, this electron has been emitted at some initial velocity by the electrode having a higher potential. Since F acts in the direction opposite to u_0, the electron is retarded or, describing the process more precisely, the electron is in a rectilinear, uniformly retarded motion, and the field is referred to as a *retarding field*. In consequence, one and the same field may be accelerating for some electrons and retarding for others, depending on the direction of the initial velocity of the electrons.

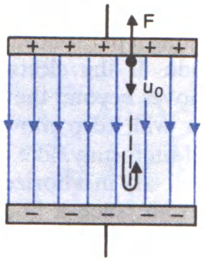


Fig. 10-2
Motion of an electron in a decelerating electric field

The force F is defined as before by Eq. (10-2), but the acceleration must be taken with a "–" sign. The kinetic energy of electrons in the presence of a retarding field is reduced because the work in moving them is done not by the field but the electrons themselves in overcoming the opposition of the field forces. Thus, an electron moving in a retarding field gives up some of its energy to the field.

If the initial energy of an electron is eV_0 and it passes in the presence of a retarding field through a potential difference V, its energy is reduced by eV. When $eV_0 > eV$, an electron will be able to traverse all of the spacing between the electrodes and to impinge upon the electrode having a lower potential. If, on the other hand, $eV_0 < eV$, an electron, on having fallen through the potential difference V_0, will have lost all of its energy, its velocity will be zero, and it will be accelerated in the reverse direction. For its travel in the reverse direction, the electron has an initial velocity of zero and moves in an accelerating field which returns to the electron the energy it lost while moving in the retarding field.

When such an electron arrives at the electrode having a higher potential, it has been accelerated to u_0. In other words, the electron moves similarly to a body which was thrown vertically upwards and allowed to fall freely, if there were no opposition from air.

The motion of an electron in a uniform transverse field. If an electron is ejected at u_0 and at right angles to the lines of force (Fig. 10-3), the field will exert on it a force F defined by Eq. (10-2) and directed towards the higher potential. In the absence of the force F, the electron would have been in a rectilinear and uniform motion at velocity u_0 due to inertia. Actually, the force F causes the electron to be in a uniformly accelerated motion at right angles to the direction of u_0. The resultant motion of the electron is along a parabola and towards the positive electrode. If the electron misses the electrode and moves beyond the field, as shown in the figure, it will keep moving by inertia rectilinearly and uniformly, like a body thrown at some initial velocity in a horizontal direction. If there were no air, such a body would have descended by gravity along a parabolic trajectory.

When an electron enters the field at an acute or an obtuse angle, it likewise traces out a parabola, and it may either hit one of the electrodes or move beyond the field.

As has been shown, an electric field always changes the kinetic energy and velocity of an electron one way or another. Thus, the electron and the electric field always exchange energy. If the initial velocity of an electron makes an angle with the lines of force rather than points along the flux, the electric field will additionally bend the travel path of the electron. Now the velocity of an electron striking the electrode will be determined solely by its initial velocity and the potential difference through which it has fallen between the terminal points of the path, irrespective of the potentials existing at intermediate points.

10-2 The Motion of Electrons in a Nonuniform Electric Field

Nonuniform electric fields have widely differing and often very complicated patterns. In them, field intensity varies from point to point in a multitude of ways, and the lines of force are usually curved in all possible manners. All of these factors complicate the study of electron motion in such fields.

The simplest case is the radial nonuniform field formed between cylindrical electrodes and often encountered in electron tubes (Fig. 10-4*a*).

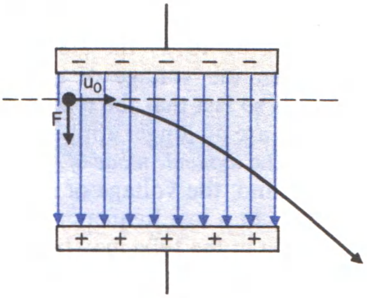


Fig. 10-3

Motion of an electron in a uniform transverse electric field

If the initial velocity of an electron ejected by the surface of the inner electrode is directed along the lines of force, the electron will move rectilinearly and at some acceleration along a radius. On moving away from the inner electrode, however, the field intensity (the flux density or the number of lines of force per unit area) and the force exerted on the electron are progressively reduced, and, in consequence, the acceleration is also reduced.

In the more general case, the lines of force in a uniform field are curves. If this is an accelerating field (Fig. 10-4*b*), an electron starting out at an initial velocity u_0 will travel along a path curved to the same degree as the lines of force. If an electron had no mass and, in consequence, no inertia, it would travel along the lines of force. But an electron has both mass and inertia, and so it tends to move along straight lines and at the velocity it acquired during the previous travel. However, the force exerted by the field on an electron is tangent to the lines of force and, since the lines of force are curved, it makes an angle with the velocity vector of the electron. Therefore, the electron path is bent, but lags behind in this bending from the lines of force owing to the electron inertia.

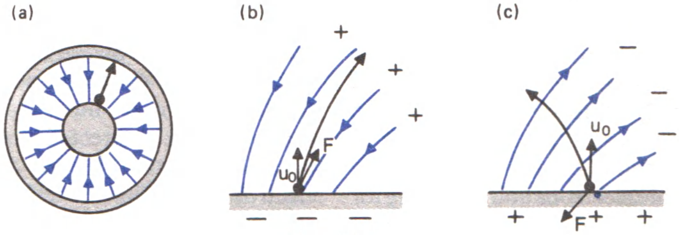


Fig. 10-4

Motion of an electron in a nonuniform electric field

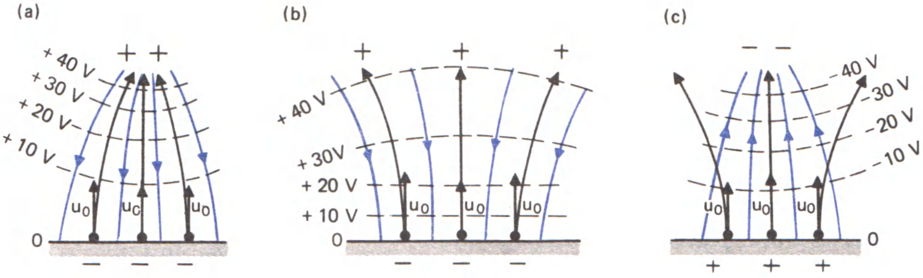


Fig. 10-5

Electron-beam focusing and spread in a nonuniform electric field

In a retarding nonuniform field with curved lines of force (Fig. 10-4*c*), the force exerted on an electron by the field likewise bends the path and changes the velocity of electrons. However, the bending is away from the lines of force, and the velocity of the electron is reduced because it moves towards points at lower potentials.

Let us examine the motion of an electron stream in a nonuniform field, neglecting the interaction of the electrons in the stream, or beam, for simplicity. The motion of an electron beam in an accelerating nonuniform field is shown in Fig. 10-5*a*. When the lines of force converge in the direction of electron travel, we have what may be called a *converging field* – its intensity is increased towards the point of convergence. Let a beam of electrons whose velocities are parallel enter this field. For simplicity, the figure shows only the middle and outer electrons of the beam. Obviously, their paths are bent in the same direction as the lines of force, and only the middle electron keeps moving rectilinearly along the central line of force. As a result, the electrons are brought closer together – the electron beam is said to be *focused. Electron-beam focusing* is not unlike light beam focusing by a collecting lens. In addition, the electron beam is accelerated.

When the lines of force diverge in the direction of electron-beam travel (Fig. 10-5*b*), we have what is called a *diverging* (or *divergent*) *field.* In such a field, the electron paths move away from one another, and the electron beam is spread sidewise. Therefore, a diverging accelerating field acts as a defocusing lens for the electron beam.

If we have a converging retarding field (Fig. 10-15*c*), the electron beam will be defocused and retarded. Conversely, a diverging retarding field will focus the electron beam. Electron-beam focusing and defocusing are widely used in many electron tubes.

Electron optics deals with many other cases of electron motion in a nonuniform field. As a rule, the field is then represented with the aid of equipotential surfaces* or, rather, by the lines along which these surfaces intersect the plane of the drawing, represented by the dashed lines in Fig. 10-5. In this figure, equipotential surfaces are labelled every 10 V. Where the lines of force come closer together, the field intensity is higher, and the equipotential surfaces are spaced closer apart. The bending of electron paths is shown by a kink where they pass through an equipotential surface. This change of direction is similar to the refraction experienced by light beams.

10-3 The Motion of Electrons in a Uniform Magnetic Field

Some electron devices depend for their operation on the motion of electrons in a magnetic field. We will investigate the motion of electrons in a uniform magnetic field because the theory explaining the motion of electrons in a nonuniform magnetic field is too complicated to be taken up in a text like this. When a magnetic field is only slightly nonuniform or when precise quantitative results need not be sought for, we may well use the laws formulated for the motion of an electron in a uniform field.

Let an electron enter a uniform magnetic field at an initial velocity u_0 directed at right angles to the magnetic lines of force (the magnetic flux) (Fig. 10-6). In the circumstances, the electron is acted upon by what is called the *Lorentz force* F which is perpendicular to the velocity vector

* An equipotential surface is at right angles to the lines of force, and all of its points are at the same potential.

$\vec{u}_0$ and the magnetic flux density (or magnetic induction) vector $\vec{B}$:

$$\vec{F} = e\vec{u}_0\vec{B} \tag{10-13}$$

As is seen, at $u_0 = 0$, the Lorentz force is zero, which implies that the magnetic field exerts no force on a stationary electron.

The Lorentz force bends the electron path into an arc of a circle. Because $\vec{F}$ is at right angles to $\vec{u}_0$, it does no work. The energy and velocity of the electron remain unchanged, but the direction of its velocity does change. As will be recalled, a body moves in a circle (rotates) at a constant velocity owing to the action of a centripetal force (that is, a force directed towards the centre of a circle), or the force F.

Several rules have been formulated by which the direction of an electron's motion in a magnetic field can conveniently be determined. One rule states: *If the observer is looking in the direction of the magnetic lines of force (the magnetic flux), the electron will appear moving clockwise.* The same rule may be re-stated like this: *The rotation of an electron is coincident with the rotation of a screw advancing in the direction of the magnetic lines of force.*

Let us determine the radius r of the circle described by a moving electron. To this end, we can use an expression for a centripetal force, known from mechanics:

$$F = mu_0^2/r \tag{10-14}$$

and equate it to the force F as defined by Eq. (10-13):

$$mu_0^2/r = eu_0B$$

Hence,

$$r = mu_0/eB \tag{10-15}$$

The higher the velocity u_0 of an electron, the more it tends to move in a straight line by inertia and the greater the radius of curvature of its path. On the other hand, an increase in B leads to an increase in F, the path is bent more, and the radius of curvature is reduced.

The above equation holds both for electrons and for any particles of any mass and charge. Therefore, we will examine the dependence of r on m and e. The higher the mass m, the more a charged particle tends to move in a straight line by inertia; and so the radius r becomes greater. The greater the charge e, the greater the force F, and the path is bent increasingly more so that the radius is reduced. Outside the range of the magnetic field, an electron will keep moving in a straight line. If, on the other hand, the radius of the path is small, an electron may describe closed circular paths inside the magnetic field.

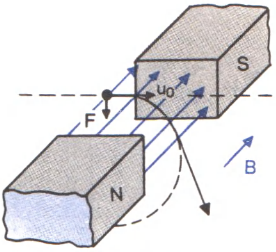


Fig. 10-6

Motion of an electron in a uniform transverse magnetic field

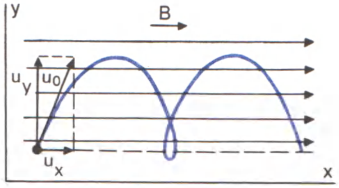


Fig. 10-7

Motion of an electron along a spiral path in a uniform magnetic field

It is an easy matter to say how an electron will interact with a magnetic field in the more general case when the electron enters the field at any angle (Fig. 10-7). Let us choose a coordinate plane such that the initial velocity vector of an electron lies in that plane, and the x-axis is in the same direction as the vector $\vec{B}$. We then resolve $\vec{u}_0$ into rectangular components $\vec{u}_x$ and $\vec{u}_y$, each running along the respective coorinate axis. The motion of an electron at velocity u_x is equivalent to a current along the lines of force, but the magnetic field does not act on such a current.

Hence, u_x experiences no changes. If an electron had only this velocity u_x, it would be moving in a straight line and at a uniform rate. In contrast, u_y is at right angles to the magnetic lines of force, and the field affects it in the same manner as it does in the case we discussed earlier (Fig. 10-6). Having only the velocity u_y, an electron would describe a circular path in a plane which is at right angles to the magnetic flux.

The resultant motion of an electron is along a spiral path. Depending on the values of B, u_x and u_y, this spiral path may be a tight one or a broad one. Its radius can readily be found on substituting the value of u_y into Eq. (10-15).

It is to be stressed once more that a magnetic field only changes the direction of an electron's velocity, but does not affect its magnitude. This implies that there is no energy exchange between an electron and a magnetic field. Thus, in comparison with an electric field, a magnetic field produces a more limited action on electrons. This is the reason why a magnetic field is used more seldom than an electric field for the control of an electron beam.

Chapter Eleven

The Basic Structure and Operation of Electron Tubes

11-1 A General Outline and Classification of Electron Tubes

Electron tubes can be used to transform electrical quantities, such as current or voltage, in waveform, magnitude and frequency, and also to convert radiant energy into electricity and back. Some tubes can convert an optical (visible) image into an electric current having an appropriate waveform or a current into a visible image (such as in TV and CRO tubes). Tubes provide a means for controlling electric, light and other quantities continuously or stepwise at a high or a low rate and with a low energy input into the control process itself, that is, without an appreciable impairment in overall efficiency. Owing to their low inertia, electron tubes can effectively operate over a huge frequency range extending from zero to 10^{12} Hz.

The above advantages of electron tubes have made them well suited for rectification, amplification, signal generation, frequency conversion, oscilloscopy of electrical and nonelectrical quantities, automatic process control, TV transmission and reception, various measurements, and many other applications.

An *electron tube* refers to a device in which the work space bounded by a gas-tight envelope is highly rarefied or filled with a suitable medium (a vapour or a gas) and which depends for its operation on the electric phenomena that take place in a vacuum or a gas.

A *vacuum* refers to a state of a gas, notably air, under a pressure which is lower than atmospheric. As applied to electron tubes, the term 'vacuum' is defined from the manner in which electrons move in a device. If electrons are free to move in the work space without colliding with the gas molecules remaining after the envelope has been pumped out, we will have a vacuum. If, on the other hand, electrons do collide with gas molecules, we will have a rarefied gas.

Electron tubes may be classed into *vacuum tubes* in which the current flow is constituted solely by electrons moving in a vacuum, and *gas-filled* (or simply *gas*) *tubes* which depend for their operation on an electric discharge in a gas (or a vapour).

In vacuum tubes, ionization is practically non-existent, and the work space is pumped out (evacuated or rarefied) to a pressure of less than 100 μPa which is classed as a high (or hard) vacuum.

In gas-filled tubes, the pressure may be 1.33×10^{-1} Pa (10^{-3} mm Hg) and higher. A substantial proportion of the moving electrons collide with the gas molecules and ionize them.

One way to classify vacuum tubes is according to the function(s) they are designed to perform. Thus we have transmitting, amplifying, rectifying, frequency-changing, detector, instrument and other tubes. Most of them are intended for continuous-wave (CW) working, but there are also tubes specifically designed for pulse working. In them, current flows in the form of pulses whose width is usually a small fraction of the spacing between them.

Another way to classify tubes is by frequency into *low-frequency* (usually *audio-frequency* or *a.f.*), *high-frequency* (usually *radio-frequency* or *r.f.*), and *microwave tubes.*

In all electron tubes, the electron stream or beam can be controlled by causing an electric or a magnetic field to act upon it.

Vacuum tubes having two diodes, a cathode and an anode, are called *diodes.* Rectifying diodes for heavy duty have the special name of *kenotrons.* Tubes fitted with control electrodes in the form of an open structure such as a mesh or a plate with a hole in it (commonly called a *grid*) may have from three to eight electrodes and are respectively called *triodes*, *tetrodes*, *pentodes*, *hexodes*, *heptodes*, and *octodes*. Tubes with two or more grids are sometimes referred to as *multi-grid tubes*. If a tube has several sets of electrodes, each with an electron beam of its own, they are called *multi-unit tubes* (such as dual diodes, dual triodes, triode-pentodes, dual diode-pentodes, and so on).

The basic types of gas-filled tubes are *stabilizer diodes*, *thyratrons*, *character-indicator tubes*, *mercury-arc tubes*, etc.

There is a large group of *cathode-ray tubes* which include *kinescopes* (or *TV picture tubes*), *TV camera tubes*, *CRO* and *storage tubes*, *electron-optical image converters*, *beam-switching tubes*, *radar* and *solar CRTs* (or *indicator tubes*), etc.

The group of *phototubes* includes *vacuum phototubes*, *gas phototubes*, and *electron multipliers.*

A special place is occupied by *X-ray tubes*, *particle counter tubes*, and some other special-purpose devices.

Electron tubes may further be classified according to the type of cathode used (hot or cold), the envelope material (glass, metal, ceramic, or combination), and the form of cooling (natural or radiant and forced by air, water, or evaporative).

11-2 The Basic Structure and Operation of the Diode

Diodes are primarily designed to rectify alternating current. Sometimes diodes may be used to generate noise (that is, currents and voltages varying in a random manner), to limit pulses, etc.

A diode has two electrodes enclosed in an evacuated glass, metal, or ceramic envelope. One electrode is a *hot cathode* which serves to emit electrons. The other electrode, an *anode*, serves to collect the electrons emitted by the cathode. The cathode and anode of a vacuum diode are similar, respectively, to the emitter and base of a semiconductor (or crystal) diode. The anode will attract electrons if it is at a positive potential with respect to the cathode. The space between the anode and the cathode is occupied by an electric field which accelerates the electrons emitted by the cathode. Acted upon by the field, the emitted electrons stream towards the anode.

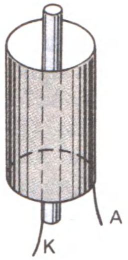


Fig. 11-1
Cylindrical electrodes of a diode

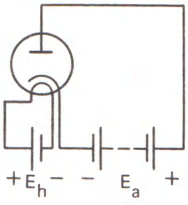


Fig. 11-2
Circuits of a diode using an indirectly heated cathode

In the simplest case, the cathode is a wire heated by current to a temperature sufficient for electrons to free themselves from its surface. This is what is known as the *directly heated cathode.* Wide use is made of *indirectly heated cathodes* in which case the cathode proper is a metal cylinder given a coat of active material that emits electrons. The cylinder encloses the *heater* which is a wire raised to incandenscence by current. In the most common type of vacuum diode, the anode is a cylinder (Fig. 11-1) whose axis is the cathode.

The circuits of a diode using an indirectly-heated (or *heater*) cathode are shown in Fig. 11-2. The anode circuit is the main circuit. It includes an anode supply source E_a and the anode-to-cathode space.

All electrons emitted by the cathode constitute what is known as the *emission current*

$$I_e = Ne \qquad (11\text{-}1)$$

where N is the number of electrons emitted per second, and e is the electron charge.

In the anode-to-cathode space, the electrons form a negative space charge which impedes the travel of electrons towards the anode. When the potential at the anode is not positive enough, not all of the emitted electrons are able to overcome the effect of the space charge, and some of them fall back to the cathode.

The electrons that do not come back to the cathode make up the *cathode current* designated as I_k or i_k:

$$i_k = ne < I_e \tag{11-2}$$

where n is the number of electrons that are emitted by the cathode every second and do not fall back to the cathode.

The higher the potential at the anode, the greater the number of electrons that are able to overcome the effect of the space charge and to reach the anode, that is, the greater the cathode current.

The stream of electrons moving inside the diode from cathode to anode and reaching the anode constitutes the *anode current* of the diode. It flows in the anode circuit and is given the symbol I_a or i_a. In a vacuum diode, the cathode current is always equal to its anode current

$$i_a = i_k \tag{11-3}$$

The anode current is the principal current of a vacuum tube. The electrons that constitute it move from cathode to anode inside the tube and from the anode to the "+" terminal of the anode supply source outside the tube, then within the source, and finally from the "+" terminal of the supply source to the cathode of the tube.

A change in the positive potential at the anode brings about a change in the cathode current and in the equal anode current. This is the essence of the electrostatic control of anode current.

Should the anode be negative with respect to the cathode, the field between them will retard the electrons escaping from the cathode and drive them back. In such a case, both the cathode and anode currents will be equal to zero.

The most important property of a diode is that it conducts current in one direction only. Electrons are able to move only from the hot cathode to the anode held at a positive potential with respect to the cathode. Should the anode be negative with respect to the cathode, the diode will be rendered non-conducting, and the anode circuit will in effect be opened. A negative anode repels electrons, but it is not raised to incandescence and cannot emit any electrons itself. Thus a diode has the property of *unidirectional conduction* and is able, similarly to a semiconductor diode, to rectify alternating current. In contrast to semiconductor diodes, however, practically no current is flowing in a vacuum diode when a reverse voltage is applied to it. In a.c. rectification, the anode supply source has an alternating emf component.

Anode current is a fraction of a milliampere in low-power diodes such as are used in radio receivers or instruments. In high-power diodes used in heavy-duty rectifiers (kenotrons), anode current may be several hundred milliamperes or even greater.

Anode current will flow subject to two conditions: (1) the cathode should be raised to a temperature sufficient for electron emission to take place and (2) the anode must be positive with respect to the cathode.

The difference in potential between the anode and the cathode is called the *anode voltage* and is given the symbol V_a or v_a. A positive anode voltage sets up an accelerating electric field which drives electrons from cathode to anode.

In practical circuits where the anode circuit contains a load across which some of the anode supply voltage is dropped, the anode voltage is smaller than E_a. It is important to differentiate between the two voltages. Confusion sometimes arises when the anode supply voltage E_a is erroneously called the anode voltage. They are equal only when the terminals of the anode supply source are connected to the anode and cathode of a tube directly (Fig. 11-2). The positive anode voltage in low-power diodes is a fraction of a volt or several volts. In medium-power rectifying diodes, it is tens of volts, and in kenotrons it is hundreds of volts or even greater.

By convention, the cathode potential is taken as datum or reference (zero potential) because electrons set out on their travel to the anode at the cathode. The potential of any other electrode in a tube is determined relative to the cathode. In the case of a direct-heated cathode, the point of zero potential is assumed to be at the "−" terminal of the filament voltage supply source.

The other circuit of a diode is the *filament circuit* (in the case of a directly heated cathode) or the *heater circuit* (in the case of an indirectly heated cathode). It contains the filament (or heater) supply voltage source, symbolized as E_f or E_h, as is appropriate. The filament (or heater) current is designated as I_f (or I_h), and the filament (or heater) voltage, that is, the voltage between the filament or heater terminals, as V_f (or V_h). The filament (heater) voltage is always low, being units or, seldom, tens of volts. The filament (heater) current is usually greater than the anode current. In low-power tubes, it is tens of milliamperes; in high-power tubes it may be as high as tens or even hundreds of amperes. If E_f is in excess of the normal value for V_f, it is usual to place a rheostat or a fixed swamping (or dropping) resistor in the filament (heater) circuit. A rheostat may also be used to adjust the filament voltage and current. The filament voltage is read by a voltmeter placed in shunt with the filament (or heater).

In many circuits, the cathode lead is connected to the metal case or chassis (Fig. 11-3). When several tubes are energized from a single filament supply voltage source, their heaters or directly heated cathodes are connected in parallel.

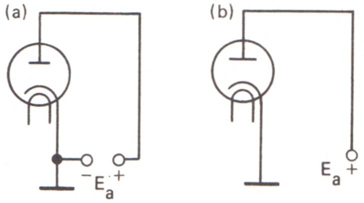


Fig. 11-3
Simplified circuits using a diode

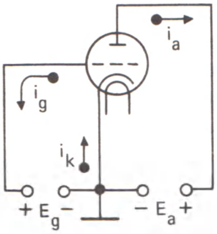


Fig. 11-4
Currents in the triode circuits

11-3 The Basic Structure and Operation of the Triode

A triode has a third electrode called the *control grid* or simply the *grid*. It is placed between the anode and the cathode. In US usage which is becoming prevalent throughout the literature of the subject, the anode of all control-type tubes to which the triode belongs is usually called the *plate* for the reason that in the early makes of tubes it was made in the form of a plane electrode.* The purpose of the grid is to provide electrostatic control of the anode, that is, plate current. A change in the grid potential with respect to the cathode brings about a change in the electric field strength and, in consequence, a change in the cathode current of the tube.

In present-day triodes. the cathode and anode are similar to the cathode and anode of a diode. The grid is most often made in the form of a mesh of fine wires. In terms of the functions they perform, the cathode, grid, and anode of a vacuum triode are similar to the emitter, base and collector of a bipolar transistor or the source, gate and drain of a FET, respectively.

All quantities associated with the grid or the grid circuit have the subscript '*g*' (which is short for 'grid').

A vacuum triode has three circuits. They are the filament circuit and the anode circuit similar to those in a diode, and also the grid circuit (Fig. 11-4). The grid circuit consists of the cathode-grid space inside the tube and the grid supply voltage source E_g. In practical circuits, some other circuit elements are included in the grid circuit.

The difference in potential between the grid and the cathode is called the *grid voltage* and is given the symbol v_g or V_g. When the grid is at a positive potential, some of the electrons are intercepted by the grid, and a grid current is flowing in the grid circuit, designated as i_g or I_g. The part of a triode consisting of the cathode, grid and cathode-grid space operates like a diode.

The principal and useful current in a triode is the anode current. It is similar to the collector current of a bipolar transistor or the drain current of a FET. The grid current, similar to the base current of a transistor, is useless and, indeed, harmful. As a rule, it is a fraction of the anode current. In many cases, measures are taken to prevent the flow of any grid current. For this to happen, the grid must be at a negative potential so as to repel electrons. Owing to the possibility of preventing the flow

* In this translation, preference is given to 'anode' for consistency in the terminology, symbols, etc.– *Translator's note.*

of grid current, the triode substantially differs from the bipolar transistor which always operates with some base current flowing.

The cathode lead carries a sum of currents, called the *cathode current*

$$i_k = i_a + i_g \tag{11-4}$$

The cathode current is analogous to the emitter current of a bipolar transistor or the source current of a FET. To repeat, the cathode current of a diode is always equal to its anode current; in a triode the anode current is equal to the cathode current only when $v_g < 0$ because then $i_g = 0$.

Similarly to the diode, the vacuum triode has the property of unidirectional conduction. However, it is not warranted to use triodes as rectifiers because diodes are simpler in design and less expensive. Since anode current can be controlled by varying grid voltage, the principal use for triodes is to *amplify signals*. Triodes are also used to *generate signals at various frequencies*. In most cases, the operation of triodes in oscillators and other special circuits reduces in the final analysis to amplification.

11-4 Electron Emission

The principal electrode of any electron device is its cathode because it emits electrons.

Electron emission refers to the liberation of electrons from the surface of a solid or liquid into a vacuum or a gas. The minimum energy that must be imparted to an electron for its escape from an emitting material at absolute zero temperature is called the *electronic work function* of that material. It varies from one metal to another and is a few electron-volts in magnitude. The greater the work function, the harder it is to bring about electron emission. The work function is smaller for metals in which the interatomic spacing is greater in comparison with other materials. The metals that have a relatively low work function include alkali and alkali-earth metals, such as cesium, barium and calcium.

Other substances present as impurities on the surface of an emitting metal markedly affect its work function. If the surface of an emitter has a coat of a material whose atoms donate electrons to the host material, electron emission will be enhanced. Such materials are called *emission activators*. Their effect consists in that the atoms donating some of their electrons are turned into positive ions. As a result, an electrostatic field is set up between the ion layer and the host metal (Fig. 11-5). This field accelerates the electrons tending to escape from the metal, and its electronic work function is reduced.

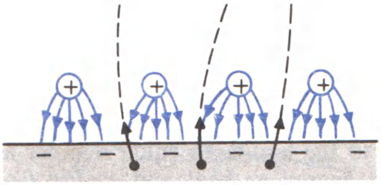


Fig. 11-5

Electric field between a metal and the positive ions of the activator

The field between the activator film and the base metal is analogous to the field in a capacitor whose plates are in the form of metal meshes (grids). In a capacitor an electrostatic field exists only between the plates, so when an electron finds its way between the wires of the negatively charged plate and into the field, it will be accelerated and leave the field through an opening in the positively charged plate.

The work function can also be reduced by giving the cathode a coat of alkali or alkali-earth metal oxides.

Electrons can be emitted in many ways.

Thermionic emission. This form of electron emission results when an electron emitting body is heated. It is widely used in vacuum tubes. As the temperature of the cathode is raised, the conduction electrons in the material acquire a greater amount of energy, and this increase may prove sufficient for them to escape from the surface of the material. If the emitted electrons are not removed from the emitter surface by an accelerating field, they form an *electron*, or *space charge, cloud* near the emitter. The electrons within the space charge differ in energy, with the most of them possessing some average energy (Fig. 11-6). The average energy usually is a few tenths of an electron-volt.

The electron space charge cloud is in a state of dynamic equilibrium: as more electrons escape from the surface of the hot material, an equal number from among those that escaped earlier return to the emitter. Thermionic emission is not unlike the evaporation of a liquid in a confined volume. There is always some saturated vapour above the liquid. The molecules in the vapour differ in energy, and the greatest number of molecules possess some average energy. The saturated vapour is in a state of dynamic equilibrium with the liquid: as some molecules return to the liquid, an equal number acquire enough energy as heat and escape from the liquid. Starting at a certain definite temperature, any further heating will cause a sudden rise in thermionic emission.

In tubes using a hot activated (say, oxide) cathode, thermionic emission can be boosted by an external accelerating field. This is the essence of the *Schottky effect*. If the cathode were not raised to incandescence, no emission would take place. With a hot electrode and in the presence of an external accelerating field, many more electrons escape from the cathode that would remain bound in the absence of such a field. Very many electrons escape from hot oxide and other activated cathodes when a strong electric field is applied instantaneously. This emission in the form of short bursts is utilized in some vacuum and gas-filled tubes.

Field (autoelectronic or cold) emission. In this form of emission, electrons are pulled out of the cathode against the surface forces by a very strong electrostatic field. In this author's opinion, the name 'cold emission' is a misnomer because all forms of emission, except thermionic emission, might be termed 'cold'.

At room temperature, electrons will escape from the surface of metals if the external electric field has an intensity of at least 10^5-10^6 V cm^{-1}.

Field emission can substantially be enhanced from a rough surface for the reason that the applied field is then concentrated at the microscopic asperities on the surface. A coat of an activating (notably oxide) material also serves to boost field emission. Apart from the reduction in the work function of the base material due to the oxide film, some of the increase in the emission is due to the penetration of the external field into the semiconductor oxide layer and the surface irregularities of the oxide.

Secondary electron emission. This form of emission takes place because the electrons produced by the cathode may strike another electrode and eject more electrons by collision. The impinging electrons are called *primary electrons*. They penetrate the surface layer and give up their energy to electrons in that layer. Some of the latter electrons may acquire enough energy for them to escape from the material. These are *secondary electrons*. Secondary electron emission usually takes place when the incident primary electrons have energies of the order of 10-15 eV and higher. The total energy of the incident primary electrons is often sufficient to liberate several secondary electrons per incident particle.

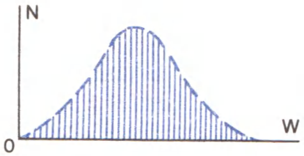


Fig. 11-6
Energy distribution of emitted electrons

The magnitude of secondary emission can be characterized in terms of the *secondary emission ratio* defined as the number of secondary electrons emitted per incident particle

$$\sigma = n_2/n_1 \tag{11-5}$$

where σ = secondary emission ratio
n_1 = number of incident primary electrons
n_2 = number of emitted secondary electrons

The secondary emission ratio σ may be smaller or greater than unity, depending on the emitter material, its surface texture, the energy of incident primary electrons, the angle of incidence, and some other factors. For pure metals its maximum value is anywhere between 0.5 and 1.8. With a coat of an activator, the secondary emission ratio may be 10 and more. Where an enhanced secondary emission is essential, resort is made to various alloys, such as an alloy of magnesium and silver, aluminium and copper, beryllium and copper, etc. Their secondary emission ratio may be in the range 2-12 or greater, and the emission is more consistent than from other materials. Secondary emission may also occur from semiconductors and dielectrics.

It is to be noted that there is no direct relationship between the secondary emission ratio and the work function. The overriding factor in secondary emission is the energy imparted by primary to secondary electrons and the ability of secondary electrons to move towards the surface from the bulk of the emitter without a substantial loss of energy. The events involved in secondary emission take place within the surface layer of the material and depend on its atomic and molecular structure.

Figure 11-7 is a plot relating the secondary emission ratio, σ, to the energy of incident primary electrons, W_1. At W_1 less than 10-15 eV, no secondary emission occurs. It begins above that figure and keeps growing with increasing W_1 until it reaches a maximum following which it falls off (σ decreases). Curve *1* applies to pure metals, and curve *2* to a metal given a coat of an activator. Secondary emission is usually a maximum when W_1 is equal to several hundred electron-volts. The reduction in secondary emission at higher values of W_1 is explained by the fact that incident primary electrons now move deeper into the emitting material and impart some of their energy to electrons lying farther away from the surface. The excited electrons then transfer this energy to other electrons and are not able to reach the surface with an energy sufficient for their escape. In a similar fashion, a stone falling into the water at a low velocity splashes it about while the same stone falling at a higher velocity will sink vithout causing any splashes.

Secondary electrons escape in various directions and at different energies. If they are not withdrawn by an accelerating field, then form a space charge at the surface of the emitting material. The energy of most secondary electrons is higher than that of thermal electrons.

Figure 11-8 shows the energy distribution of secondary electrons for some metal when the energy of incident primary electrons is $W_1 = 150$ eV. Most secondary electrons escaping from the emitter have energies in the range from 0 to 50 eV. The greater proportion has an energy of about 10 eV. Also, there is a marked number of secondary electrons whose energy is nearly equal to that of primary electrons. They are believed to be reflected electrons, although it is not unlikely that some primary electrons give up all of their energy to secondary electrons at the surface of the emitting material. Such electrons may escape without any loss of energy, that is, with an energy equal to W_1. A rise in the energy of primary electrons leads to an increase in the number of secondary electrons of low energy. This checks with the idea that primary electrons move deeper into the material so that the secondary electrons reaching the surface lose more energy in the process.

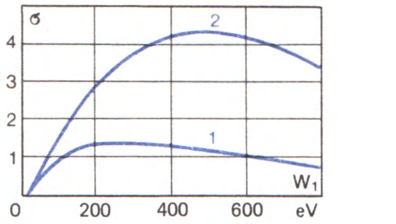


Fig. 11-7

Secondary-emission ratio as a function of primary-electron energy

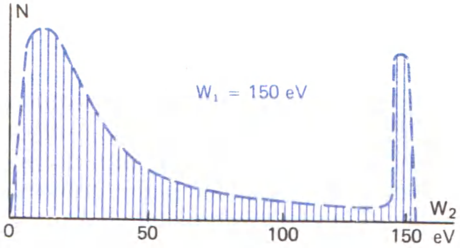


Fig. 11-8

Energy distribution of secondary electrons

For many years secondary electron emission remained unutilized for lack of stability. Later, secondary-emission cathodes were fabricated from alloys capable of a sustained performance, and vacuum tubes using secondary electron emission became a reality.

Electron emission due to bombardment by heavy particles. This form of emission has much in common with secondary electron emission. In most cases, electrons are knocked out of an emitter bombarded by ions. The extent of emission is stated in terms of the *emission yield* δ, defined as the ratio between the number of ejected electrons, n_e, and the number of incident ions, n_i:

$$\delta = n_e/n_i \qquad (11\text{-}6)$$

The value of total yield depends on the material bombarded by ions, the mass and energy of incident ions, the condition of the bombarded surface, the presence (or otherwise) of an activating coat, the angle of incidence, and some other factors. As a rule, the yield is less than unity for metals, but it may be greater than unity for semiconductors and thin dielectric layers. The yield first rises with increasing energy of incident ions, then falls off similarly to the secondary emission ratio. The least energy of incident ions required for electrons to be ejected from the emitter is tens of electron-volts. A higher yield is obtained in the presence of an activating coat. For most ejected electrons the energy is not more than 1 to 3 eV, although some of them may have energies as high as tens of electron-volts.

Electron ejection due to ion bombardment is the principal form of emission in glow-discharge gas tubes, such as gas-filled stabilizing diodes, neon tubes, and some other devices (see Chap. 17).

There is a further form of electron emission known as *photoemission*. It is examined in detail in Chap. 22.

11-5 The Parameters of Thermionic Cathodes

Thermionic cathodes are to meet a number of requirements. For one thing, a thermionic cathode must be durable and sustain emission at the least possible value of power dissipated by the filament (or heater). For another, the emission must be consistent (stable). The cathode surface ought not to be destroyed by ion bombardment. (Even in a high vacuum there is a number of positive ions which move towards the cathode with an acceleration. The higher the anode voltage, the stronger the impact of these ions on the cathode.)

If a cathode is to be used properly, it is important to know its filament (or heater) voltage V_f and its filament (or heater) current I_f.

The performance of a cathode can be stated in terms of its efficiency, *H*, or the value of emission current per watt of filament (heater) power:

$$H = I_e/P_f \qquad (11\text{-}7)$$

Most state-of-the-art cathodes, when operated continuously, have an efficiency of units to hundreds of milliamperes per watt.

The parameters of a cathode also include its *operating temperature T* and the *service life t*. The operating temperature ranges from 700° to 2300°C, depending on the type of cathode. The service life is the span of time at the end of which the electron yield is reduced by 10%. Thus, a cathode is still capable of normal operation at the end of this time interval, but at a reduced level of emission. The service life of cathodes ranges from hundreds to tens of thousands of hours.

The parameters *H*, *t* and *T* are interrelated. An increase in *T* leads to a higher *H*, but to a shorter *t*. Cathodes can operate under various service conditions. For greater emission, it may be advantageous in some cases to raise the filament voltage, but this will inevitably cut down the service life. If a longer service life is essential, a reduced filament voltage may be used. This, however, entails an impairment in efficiency.

11-6 Cathodes

Cathodes may be fabricated from pure metals, the prevalent material being tungsten and only sometimes tantalum. *Pure-metal*, or *simple*, *cathodes*, are of the *directly heated* or *filamentary type*.

Tungsten is a refractory (that is, high-melting-point) metal which melts at 3300°C. The operating temperature of tungsten cathodes is in the range 2100-2300°C which corresponds to heat colours from light-yellow to white. The limit to the service life of tungsten cathodes is set by a fall in emission caused in turn by the reduction in the cathode thickness as some of the tungsten is sputtered.

The primary advantage of tungsten cathodes is their highly consistent emission. The electron yield will not fall off even after a short-duration overheating. The stability of tungsten cathodes towards ion bombardment makes them especially suited for use in heavy-duty tubes operating at high anode voltages. Tungsten cathodes are also used in electrometer tubes where a consistent emission is especially important. Also, in tubes using tungsten cathodes, the evaporated tungsten particles lodge on the inside of the tube envelope and form a layer that absorbs gases and improves the vacuum.

A major limitation of tungsten cathodes lies in their low efficiency. Because of the high temperature at which they have to be operated, they radiate a good deal of light and heat, thus wasting a sizeable proportion of input power.

In many types of cathodes, known as *coated cathodes*, the surface of a pure-metal emitter is given a coat of an activator so as to sustain a high level of emission at relatively low temperatures.

The primary advantage of such cathodes is their economy of operation. Their efficiency may be as high as tens or even hundreds of milliamperes per watt. The operating temperature of some such cathodes is 700°C. The service life is thousands or even tens of thousands of hours. Towards the end of this time interval, the electron yield falls off due to a reduction in the amount of activating impurities (say, due to their evaporation). Some cathodes in this class show an extremely high level of emission in pulse working (during time intervals as short as several microseconds) separated by longer idle intervals.

The key limitation of coated cathodes is the low stability of emission. Even a temporary overheating (filament overvoltage) may cause a drop in electron yield because elevated temperatures cause some of the activator material to evaporate. Also, coated cathodes are destroyed by ion bombardment, so it is vital to maintain a very high vacuum in the envelope. This goal is achieved through the use of suitable *getters*, alkali metals introduced into a vacuum tube during manufacture.

Coated cathodes include *atomic-film emitters* and *oxide-coated emitters*. Among the former is the *carburized thoriated-tungsten cathode*. It is a tungsten filament given a coat of thorium and containing an amount of carbon which reacts with the tungsten to form tungsten carbide (this is why the cathode is said to be carburized). The thorium film forms a monatomic layer which evaporates at a lower rate than pure tungsten. Carburized thoriated-tungsten cathodes can operate at 1600-1700°C and have $H = = 50\text{-}70$ mA W^{-1}. The active layer of such cathodes effectively stands up to ion bombardment, so they may be used at an anode voltage of up to 15 kV.

An oxide-coated, or simply oxide, cathode uses nickel or tungsten as the base metal to which a mixture of barium, calcium and strontium oxides is applied as a thin layer. The operating temperature of oxide cathodes is 700-900°C (which corresponds to a dark-red or red heat colour). The efficiency is 50-100 mA W^{-1}. The service life is up to tens of thousands of hours.

In an oxide cathode, electrons are mainly emitted by the barium. Overheating speeds up the evaporation of barium and causes a reduction in electron yield. The limit to the service life of oxide cathodes is set by the fact that the oxide coating is gradually depleted of barium atoms. For proper operation, an oxide cathode must be held in a high (hard) vacuum because the oxide coating may readily be destroyed by ion bombardment. To avoid excessive ion bombardment, the anode voltage ought not to be very high in continuous working.

The oxide coating has an appreciable resistance and so it is additionally heated by anode current. Electron emission from an oxide cathode can be enhanced by an external accelerating field which extends into the depth of the oxide coating (the *Schottky effect*). However, it is important to guard against an excessive rise in cathode current. At an excessively heavy cathode current, overheated areas appearing as luminous spots are produced on the cathode surface. In these areas, the oxide is evaporated at an elevated rate, which may sometimes be accompanied by cathode sparking – the ejection of incandescent slivers of the cathode material. Most often, sparking occurs at a very high cathode current or at an excessive anode voltage, especially right after the filament voltage has been turned on but the cathode has not yet had time to come up to its operating temperature so that there is no emission and there is no space charge to limit the strong effect of the accelerating field on the oxide coating.

Overheated areas may be produced on an oxide cathode not only by an excessive filament voltage, but also when the filament voltage is too low. When this happens, a directly heated (filamentary) cathode may burn as the metal near a hot spot melts. This can be explained by the properties of oxide cathodes:

(1) As with all semiconductors, the resistance of the oxide coating falls off with rising temperature.

(2) Owing to the high resistance of the oxide coating, the heat input due to cathode current is comparable with that due to filament current.

(3) The thickness, resistance and emissivity of the oxide coating are all different at different places on the cathode. The total cathode current is distributed so that more of it is flowing through the areas having a lower resistance and a higher emissivity. In consequence, these areas are heated more, their resistance is further reduced, electron yield builds up, and the current rises again. These events occur when the filament voltage is too low and the cathode current is too high. Then the heat input due to the filament current is reduced, and a greater contribution comes from the cathode current.

Ion bombardment of the cathode also serves to produce hot spots.

It is not always that hot spots will appear at too low a filament voltage. Still, this danger does exist, and care must therefore be taken to avoid it in tubes with oxide cathodes, especially at a heavy cathode current.

When the filament voltage is maintained at its normal value and there is no overload by cathode current, oxide cathodes can last for a very long time. They are widely used in receiving, amplifying and transmitting tubes of low and medium power ratings, in CRTs, in tubes for pulse working, and many other electron devices.

In pulse working, that is, when the cathode is required to emit short pulses of current (several microseconds long) spaced apart by sizeable intervals, its emission may be many times the figure in continuous working. Now it is brought about by a strong external electric field so that it is a combination of field emission and thermionic emission. The electric field promotes the escape of a large number of electrons from the oxide layer. With time, however, this emission is reduced (Fig. 11-9). An extremely strong emission is said (not very accurately) to *poison* the oxide cathode.* Poisoning disappears if the cathode is allowed to recover or 'rest' for a sufficiently long time. Its emission will then be restored, and the cathode will be able to give a high electron yield for a short time. The point is that a sufficient number of electrons available for emission must be accumulated in the oxide layer. The duration of emission current pulses does not usually exceed 10 μs, and the spacing between the pulses when the cathode is recovering lasts hundreds of microseconds.

The efficiency of oxide cathode in pulse working may be as high as 10^4 mA W^{-1}. Cathode current pulses may be units or even tens of amperes in magnitude. The shorter the pulses, the greater the current. With short pulses, the cathode is not practically bombarded by ions, and so the anode voltage may be as high as 10-20 kV.

There are also *thoria cathodes* and *L-type cathodes.* Thoria cathodes are used in transmitting tubes and have a thoria coating applied to a tungsten or tantalum base. Or they may be fabricated by sintering powdered thoria with the filament. In continuous working, these cathodes have an efficiency of 300 to 2000 mA W^{-1}.

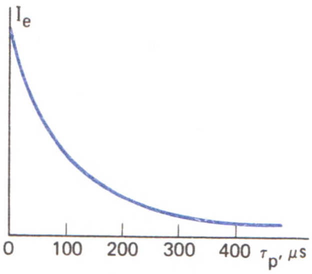


Fig. 11-9

Oxide-cathode emission as a function of anode-current pulse duration

In the L-type cathode (so called after its inventor, H. J. Lemmens), there is a reservoir or pellet of barium carbonate-strontium carbonate sealed behind a porous plug or pressed nickel powder. In fact this is the older form of the L-cathode known as the *dispenser type.* The newer design is known as the *impregnated type* in which barium oxide is dispersed throughout the pressed nickel powder or sponge. This type of cathode has a lower resistance and it tends less to sparking and burn-out.

11-7 Directly and Indirectly Heated Cathodes

A directly heated cathode is a wire or ribbon filament for which reason directly heated cathodes are alternatively called *filamentary(-type) cathodes.* The filament is usually bent in a zigzag fashion.

Among the advantages of directly heated cathodes are simple design and suitability for very low-power tubes with a filament current of 10 mA or even less. Directly heated cathodes are also used in high-power transmitting tubes and in tubes for low-power portable and mobile radio sets which draw their power from dry cells or storage batteries.

A thin-filament cathode rises to incandescence in less than 1 s after turn-on, which is an obvious convenience. A disadvantage of directly heated cathode is that parasitic pulsations are produced in the anode current when the filament is energized with an alternating current. If, for example, the heating current has a frequency of 50 Hz, the anode current will pulsate

* It would be more appropriate to say that the cathode is exhausted or fatigued.

at 50, 100, 150 Hz, etc. These pulsations corrupt and mask valid signals. In aural reception, they are heard as a low-pitched droning noise known as *a.c. hum.* There are two causes for these pulsations.

Firstly, thin cathodes experience temperature pulsations because their mass and heat capacity are very small. When the current reaches its peak value, the temperature is at a maximum; when the current crosses zero, the temperature is at a minimum (Fig. 11-10). The frequency at which the temperature of the cathode pulsates is twice the frequency of heating current. The emission and anode current pulsate at the same double frequency. This phenomenon is almost negligible in tubes with more substantial cathodes.

Secondly, the surface of the cathode is not equipotential. Its surface potential varies from one point on the surface to another, and the anode voltage corresponding to these points is likewise different. Therefore, when a directly heated cathode is energized with an alternating current, the anode voltage pulsates at the frequency of the heating current, and so does the anode current.

Another disadvantage of tubes with fine directly heated cathodes is *microphonics*, that is, the mechanical translation of vibration or shock into an electrical signal by a vacuum tube. The point is that external shocks cause the cathode to vibrate so that the interelectrode spacings are changed. In turn, this leads to variations in the anode current. In aural reception, these variations are perceived as an electric signal. Microphonics is also responsible for *acoustic feedback* (also known as *acoustic regeneration*). In this case, sound waves from a speaker cause the tube to vibrate so that the anode current experiences variations which, upon amplification, reach the speaker. The resultant sound waves again act on the tube, and the process is repeated all over again, giving rise to undamped sound oscillations which bury or mask valid signals. In high-power tubes using substantial cathodes, microphonics is negligible.

An indirectly heated cathode consists of a nickel tube which is coated with an emitting oxide coat and encloses a coiled tungsten heater as shown in Fig. 11-11*a* (which is the reason why such cathodes are alternatively called the *heater type*). The heater is insulated from the cathode proper by an aluminium oxide coating.

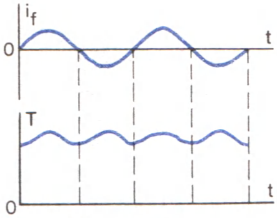


Fig. 11-10

Pulsations in the temperature of a directly heated cathode operating on alternating current

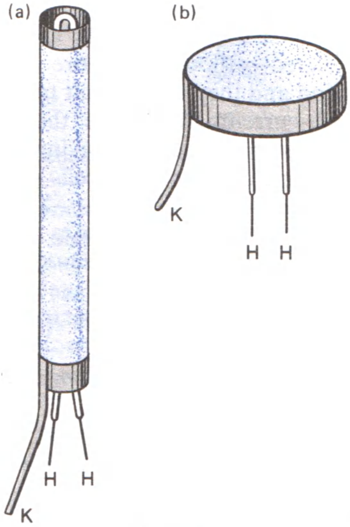


Fig. 11-11

Indirectly heated cathodes

Long heaters are usually shaped in a zigzag pattern or coiled into a helix. In some tubes, the indirectly heated cathode is a short cylinder with its top covered by an oxide (Fig. 11-11*b*). The cylinder encloses a heater insulated by aluminium oxide.

A major advantage of indirectly heated cathodes is freedom from pulsations in the anode current when they are heated with an alternating current. Temperature variations are practically nonexistent because the cathodes have a larger mass and, in consequence, a greater heat capacity than directly heated cathodes. It takes an indirectly heated cathode a few seconds to come up to full heat from the instant when the heating current is turned on and to cool completely from the instant when the heating current is turned off. At a supply frequency of 50 Hz, a quarter of a cycle lasts just

0.005 s – a time interval too short for the cathode temperature to change appreciably, so the emission current does not pulsate.

Indirectly heated cathodes have an equipotential surface. There is no voltage drop along the cathode due to the heating current. The anode voltage is the same at any point on the cathode surface. It does not pulsate when the filament voltage varies.

Another advantage of indirectly heated cathodes is the negligible effect of microphonics. The cathode mass is rather substantial, and it can hardly be made to vibrate.

As compared with filamentary cathodes, indirectly heated cathodes are more intricate in construction and they are difficult to design for very small currents. Therefore they are less suited for low-power energy-saving tubes intended for operation of dry cells or storage batteries.

When used in communications equipment (such as transceivers) which operates intermittently and must be ready for use right after turn-on, tubes with indirectly heated cathodes have to be kept with their heater voltage 'on' all the time (this is known as *hot stand-by*). This is wasteful of power and cuts down the tubes' service life. It is inconvenient to use tubes with indirectly heated cathodes in battery-powered portable low-power transceivers. To save the power supplies in such cases, one has to turn off the heater voltage of the receiver when the transmitter is operating and to turn off the heater voltage of the transmitter when the receiver is operating. In either case, one has to wait for 10 to 20 s before the cathodes come up to their operating temperature, and this introduces a delay in communication.

The hot aluminium-oxide insulation between the cathode and the heater cannot stand up to elevated voltages. The limit of voltage between cathode and heater is usually set at 100 V, and only for some tubes the limit is 200-300 V. In some circuits, the cathode and the heater have to be maintained at different potentials. If the difference should exceed some critical value, the insulation between the cathode and heater might break down and the tube might fail. No risk of breakdown exists when the cathode is connected to one of the heater's terminals (see Fig. 12-8*b*).

11-8 Anode and Grid Types for Vacuum Tubes

The anode (plate) of a vacuum tube collects the stream of electrons emitted by the cathode. The electrons carry a substantial amount of kinetic energy which they give up on striking the anode. This bombardment raises the temperature of the anode. In addition, the temperature of the anode is raised owing to the heat radiated by the cathode itself. In a steady state, the amount of heat the anode receives must be equal to the amount of heat withdrawn from it.

It is important that the anode should not be raised above its temperature limit. Overheating may cause the anode to give up gases, and this will impair the vacuum inside the envelope. An incandescent anode emits thermal radiation which may overheat the cathode. Excessive overheating may even melt the anode.

To avoid overheating, measures are usually taken to cool it. In low-power tubes and most of the medium-power tubes, the anode is cooled by radiation (*radiation cooling*). Heat is abstracted by thermal radiation which is emitted by the anode and absorbed by the envelope in part or completely, depending on the envelope material, and raises its temperature. In turn, the envelope gives up its heat to the surroundings.

The amount of heat radiated by an anode can be increased by increasing its surface area or by treating it in a suitable manner (for example, by making it black or matted). Frequently, the anode is fitted with fins to increase its cooling surface area (Fig. 11-12). Medium- and high-power tubes often use forced air cooling. To this end, the anode connection is fitted with a suitable radiator blown over with air by a fan.

Anodes using radiation cooling should be fabricated from materials which are heat-resistant and are good thermal emitters. Importantly, an anode ought not to liberate any

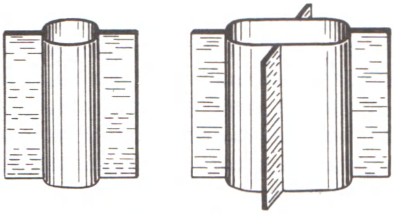


Fig. 11-12

Finned anodes for better cooling

dissolved, entrapped or occluded gases. Most often, anodes are made of nickel, nickel-plated or aluminized steel, and molybdenum. High-power tubes use anodes made of tantalum or graphite.

High-power tubes may alternatively use forced water cooling (Fig. 11-13). A forced-water-cooled anode is a copper or chromium-steel cylinder sealed to a glass envelope. The electrode leads are passed through the glass. The anode is enclosed in a jacket through which distilled water is circulated. Ordinary tap water is unsuitable for cooling purposes because it forms scale on the anode surface, thus impairing heat abstraction. Cooling water is supplied by a pump and circulated in a closed system. Heated by the anode, the water is then cooled and returned to the anode so as to absorb more heat, and so on. The temperature of forced-water-cooled anodes does not exceed 100-120°C. Water-cooling systems are fitted with automatic controls that regulate the flow rate of cooling water and will turn off the anode supply voltage, should the flow of water be interrupted for one reason or another.

Some tubes employ what is known as *evaporative cooling*. A typical evaporative-cooling system consists of a tube with a specially designed anode immersed in a boiler containing distilled water. When power is dissipated at the anode, the water boils and the steam is conducted upwards through an insulating pipe to a condenser. The condensate is then gravity-fed back to the boiler, thus eliminating the pump required in a circulating water system.

When the anode of a tube is a cylinder coaxial with its cathode, its grid (or grids) is usually wound of fine wire on a heavier frame. This type of grid is shown in Fig. 11-14*a*. Sometimes, a grid is a fine-wire woven mesh such as shown in Fig. 11-14*b*. The grid for a triode using plane electrodes is shown in Fig. 11-14*c*.

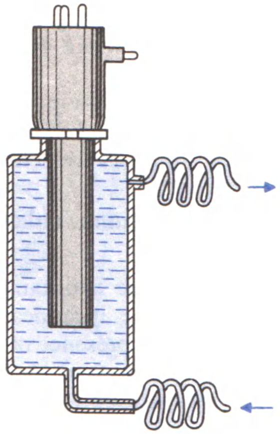


Fig. 11-13

Water-cooled anode of a high-power tube

Grids are usually fabricated from nickel, molybdenum, or their alloys. More seldom they are made of tantalum or, when the grid wires have to be very fine, of tungsten. The performance of a tube would be impaired if the grid were allowed to grow hot due to radiation from the incandescent cathode because it would then emit thermionic electrons of its own. To avoid this, grid wires are given a coat of a metal which has a greater electronic work function, such as gold.

It is usual to place the grid of a tube closer to its cathode for better control of the electron

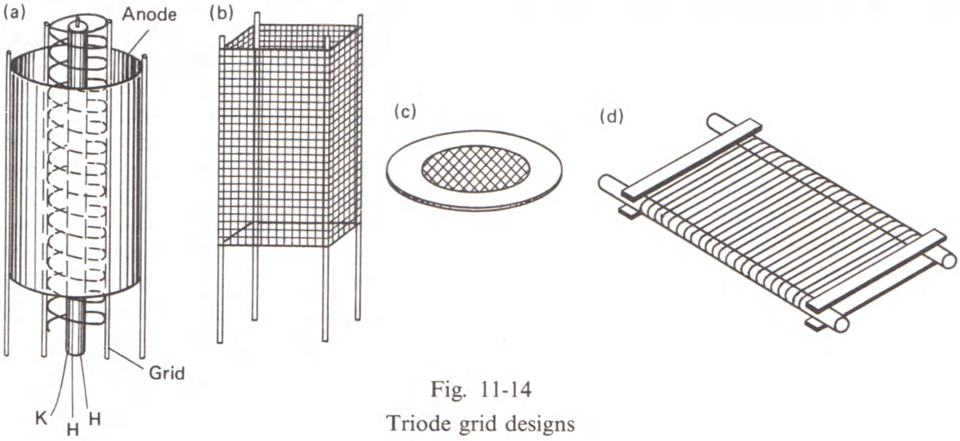


Fig. 11-14

Triode grid designs

stream. This is done for example in tubes using a grid wound of tungsten wire 8-10 μm on a frame (Fig. 11-14*d*).

For better cooling, low-power tubes use blackened grid wires, and high-power tubes have their grid supports welded to radiators.

11-9 Vacuum in Vacuum Tubes. Tube Envelopes

A vacuum has to be maintained in a tube above all because the incandescent cathode would burn out in the presence of air. Also, the gas molecules would prevent the free travel of electrons. In high-vacuum (or hard) tubes, the pressure is less than 100 μPa. When the vacuum is not hard enough, moving electrons collide with gas molecules and turn them into positive ions which bombard and damage the cathode. Gas ionization also adds to the transit-time effects and instability in a tube's operation and produces additional noise.* Thus, a gassy or 'soft' tube, unless it is intended to be such, cannot do its job properly.

During tube manufacture, the envelope is first evacuated to a rough vacuum of 1.33-0.0133 Pa (10^{-2}-10^{-4} mm Hg) by roughing-down vacuum pumps. Following that, the envelope is exhausted by high-vacuum diffusion pumps. Then the metal parts inside the tube are degassed by heating them with r. f. power or, for metal tubes, with a flame. Thus heated, the metal parts give up their occluded or absorbed gas which is then evacuated by a pump.

The vacuum in a tube is further improved by placing inside it the *getter* – a material that has a strong chemical affinity for any residual gases, such as a piece of magnesium or barium or their alloy with aluminium. The getter structure is first degassed with r. f. power or, in certain cases, by a flame or by direct electric heating. Then, just before the exhaust tubing is to be sealed off, the getter is 'flashed' by heating it to a temperature at which a chemical reaction takes place; a free getter metal is produced and deposited on part of the tube envelope as a mirror-like metallic film which is silvery white in the case of magnesium or brownish-black in the case of barium. After the exhaust tubing has been sealed off, this film acts as a chemical pump that will adsorb or react with any gases evolved during the life of the tube, thus maintaining a high vacuum.

Tube envelopes can be made of glass, metal, ceramic materials, or their combinations. The size of an envelope depends on the power rating of the tube for which it is intended. Thermal radiation emitted by the electrodes pass in the part through the envelope, while the remainder is absorbed by the envelope. To avoid a heat build-up on the envelope, measures are taken to enlarge its surface area. Glass envelopes are most common, but ceramic envelopes are more heat-resistant and robust mechanically.

Metal (steel) envelopes are very strong and effectively shield the electrodes against extraneous electric and magnetic fields. Unfortunately, they become very hot, and this leads to electrode overheating. Tubes enclosed in metal envelopes have a greater number of metal-to-glass seals which are likely to pass air. Of late, the manufacture of metal envelopes has been discontinued, but such tubes may be encountered in the older makes of radio and electronic equipment.

11-10 Electrode Mounting and Electrode Leads

In the older makes of tubes, the electrode system, technically known as the *cage assembly*, is welded to a glass *stem* which is the shape of a tube flattened at one end (Fig. 11-15*a*). Sealed into this flattened end are stem leads made of metal which expands to the same extent on heating as glass, such as alloys of nickel, cobalt and iron (known as Covar and Fernico). The stem leads are welded to the electrodes themselves or to thin metal connectors called *tabs*. The opposite ends of the stem leads are welded to the wires that run to the *base pins*. The stem also passes a glass exhaust tube through which air is pumped out of the envelope. Sometimes, an envelope will be pumped out through an exhaust tube at the top of the envelope.

The parts of a tube are usually held in precise alignment by means of spacers punched from thin sheets of mica or ceramic material which are placed at the top and bottom of the tube structure (Fig. 11-15*b*).

In bantam and some other tubes, the cage assembly is mounted on a small glass button which is the base of the envelope. There are leads

* Ionization serves a useful purpose in gas-filled tubes.

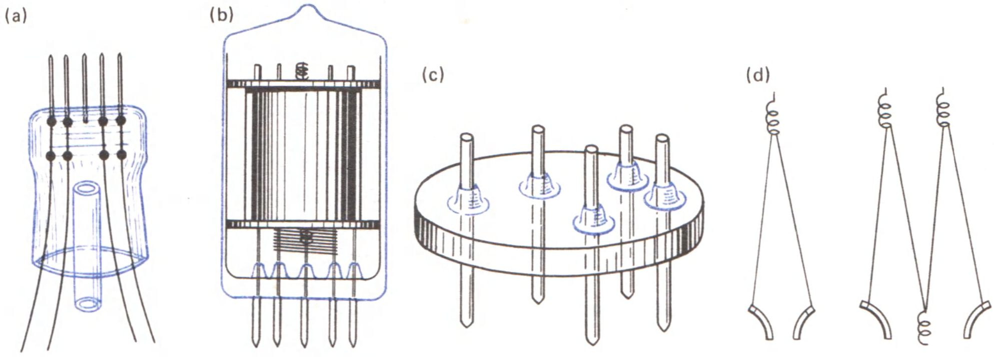


Fig. 11-15
Mounting of electrodes and leads in glass tubes

sealed into the glass button (Fig. 11-15*c*). Outside the tube, they do the job of tube pins, and inside the tube they support the electrodes (Fig. 11-16*a*). The cathode is of the directly heated type and is usually stretched taut by a spring (Fig. 11-15*d*), so that it will not sag when it expands on heating.

In metal tubes, the lower part of the envelope is welded to a steel bottom which passes an exhaust tube and has holes into which Fernico bushings are welded. Into the bushings are sealed drops of glass enclosing the leads and tabs (Fig. 11-16*a*). Use is also made of a solid glass bottom passing a glass exhaust tube. The tabs and leads are then sealed into a glass bottom (Fig. 11-16*b*).

Tubes also have some other parts for auxiliary purposes. They include tabs and shelves for the getter structure, electrostatic shields to prevent capacitive currents between the various parts of a tube or to keep extraneous electric fields out of the envelope, etc.

Special care is exercised when the parts of a tube are put together and mounted. Still, there is a good deal of spread in electric properties among the individual tubes of the same type. This comes from the differences in parts, their straining in assembly, inaccuracies in mounting, differences in emission from the cathodes of different tubes, and some other causes.

The leads are carried to the pins in a pattern usually shown on what are called *base* (or *pin*) *connection diagrams*. As an example, let us see how the leads are connected to the pins in receiving and amplifying tubes. Glass and metal tubes have a standard *octal base*, that is, one with eight pins positioned at the corners of a regular octagon (Fig. 11-17*a*). At the centre of the octagon is what is known as the *key* or the *alignment pin*. A projecting rib on the key assures that the tube is correctly inserted into its socket. As a rule, the base pins are numbered clockwise, starting from the rib on the alignment pin. The metal envelope or the electrostatic shield (if there is any inside the tube) is connected to one of the base pins. The pin connection pattern is different for different tubes, so it is important to refer to a tube's pin connection diagram before inserting it into the socket. Some

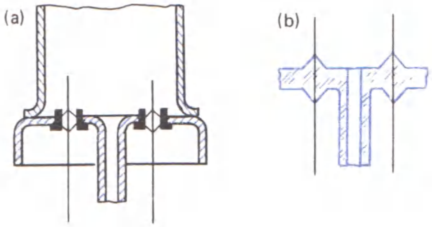


Fig. 11-16
Mounting of electrode leads in metal tubes

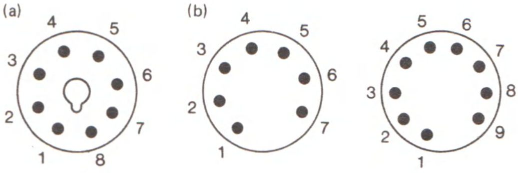


Fig. 11-17
Pin assignment in tubes

of the standard eight pins may be omitted in an actual tube if they are not connected to any electrodes.

In kenotrons intended to rectify alternating currents at several hundred volts, all electrodes have connections to the base. In kenotrons for voltages of several thousand volts, the anode and cathode leads and pins may not be placed close to each other; in them, the anode terminal is located at the top of the envelope (the anode cap).

In bantam tubes, there may be seven or nine pointed leads sealed into a glass stem and located at the corners of a regular octagon or decagon, respectively (Fig. 11-17*b*).

Very small tubes usually have no base and their leads are sealed into a small glass button. There is a colour mark or arrow at the edge of the stem to indicate the pin from which the other pins should be counted.

In high-power tubes the electrode leads are often located at different points on the envelope and at some distance from one another because a substantial voltage may exist between them.

Chapter Twelve
Diodes

12-1 Physical Processes

To begin with, we will discuss a diode having plane electrodes. The anode voltage sets up an electric field between anode and cathode. This field remains uniform so long as the cathode is not emitting electrons. In normal operation, the cathode emits a great number of electrons which produce a negative space charge in the cathode-to-anode space, which impedes the travel of electrons towards the anode. The space charge has a maximum density near the cathode. Sometimes it is said that an electron cloud is formed at the cathode (Fig. 12-1). The space charge forming between anode and cathode makes the field nonuniform.

The field may be either accelerating or retarding with regard to the emitted electrons. Accordingly, the diode can operate in any one of two basic modes. If the field is an accelerating one all along the distance from cathode to anode, all the electrons that the cathode is capable of emitting will be accelerated towards and reach the anode. None of the electrons will fall back to the cathode in these conditions, and the anode current will be maximal and equal to the emission current. The diode is then said to have reached *saturation*, or, more often, the *temperature-limited condition*, and the corresponding anode current is called the *saturation anode current* I_s:

$$I_s = I_e \qquad (12\text{-}1)$$

The principal mode of operation for vacuum tubes, however, is the *space-charge-limited region*. In this region, the field near the cathode is a retarding one for the emitted electrons. As a result, electrons starting at a low initial velocity are not able to overcome the retarding field and fall back to the cathode. The remaining electrons of a high initial velocity do not lose all of their energy in the retarding field and move on to the anode.

In this region, the anode current is smaller than the emission current

$$i_a < I_e \qquad (12\text{-}2)$$

The events taking place in a diode can be visualized by reference to potential diagrams showing the potential distribution in the anode-cathode space (Fig. 12-2). On these diagrams, the distance from the cathode is laid off as abscissa and the potential as ordinate. By convention, positive potentials are laid off down-

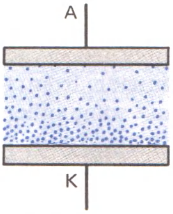


Fig. 12-1
Electron space charge in a diode

wards. The cathode is assumed to be at zero potential.

When the cathode is cold, there is no space charge, and the field is a uniform one. The potential from cathode to anode rises in proportion to the distance from a given point to the cathode (straight line *1*). If the cathode is raised to incandescence, there forms a negative space charge, and the potentials are brought down at all points except the cathode and anode because the anode voltage is set by an external source. The potential distribution curve flexes upwards (curve *2*). So long as the space charge is small, the potentials at all points remain positive (curve *2* runs below the horizontal axis), and an accelerating field exists. The potential distribution represented by curve *2* corresponds to the saturation region. As the cathode is heated progressively more, the space charge also grows, and the potentials at all points are brought down still more. The potential distribution curve also flexes upwards more, and the negative potential near the cathode may exceed in absolute value the positive potential of the accelerating field set up by the anode. The resultant potential then is negative, which is clearly shown by curve *3* running above the horizontal axis near the cathode.

At some distance x_0 from the cathode the potential decreases to its minimal value, φ_{min}. As a rule, φ_{min} is a few tenths of a volt. Over the distance from the cathode to x_0 the electric field is a retarding one for the emitted electrons, producing a *potential barrier* at the cathode. Of all emitted electrons, only those having an initial velocity sufficient for them to overcome the barrier reach the anode. Electrons with a lower initial velocity lose their energy before they reach the crest of the potential hill and fall back on the cathode. Curve *3* showing the presence of a potential barrier corresponds to the space-charge-limited region or condition. The further increase in heat input to the cathode leads to a situation represented by curve *4*: the potential barrier grows higher and moves away from the cathode.

The events we have described can be illustrated by the following mechanical analogy. Let Fig. 12-2 show a land profile and the balls rolling down from point *K* model the electrons emitted by the cathode. The balls move all at different velocities. When the ground is downsloping right at point *K* (curves *1* and *2*), all balls roll down as far as point *A*. When, however, the lay of the ground is such as shown by curve *3*, there will be a hill on the way of the balls, and only those of them that have a sufficient initial velocity will roll over the hump, and those having a lower velocity will roll back towards point *K*. This mechanical analogy explains why we have chosen the downward direction for positive potentials.

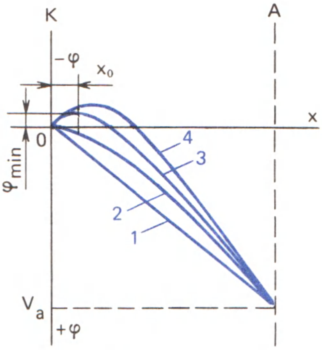


Fig. 12-2

Potential diagrams of a diode with the anode voltage held constant at several values of filament voltage

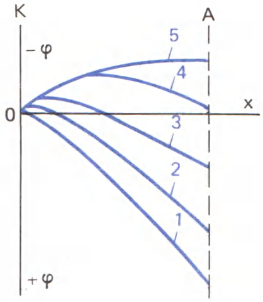


Fig. 12-3

Potential diagrams of a diode for a constant filament voltage and at several values of anode voltage

Figure 12-3 shows potential diagrams for several values of anode voltage at a constant filament voltage. At some definite voltage, the tube moves into the saturation region (curve *1*); at a lower voltage it is in the space-charge-limited region (curve *2*). At a still lower voltage (curve) the potential barrier grows higher. At $V_a = 0$, the situation is represented by curve *4*. For V_a to be equal to zero, the anode must be short-circuited to the cathode. Then electrons will form a space charge in the cathode-anode space and the potential barrier will grow still higher. Electrons starting at a high initial velocity will be able to overcome this barrier and reach the anode. Such electrons are few in number, and the initial anode current, I_0, existing at $V_a = 0$ is very small.

Curve *5* represents the situation when the anode circuit is open, that is, the anode is not connected to anything. Just as the anode circuit is opened, the anode is at zero potential as shown by curve *4*. In the circumstances, only electrons having high velocity can reach the anode and charge it negatively. The right-hand end of the diagram is shifted upwards (curve *5*), the potential barrier grows higher, and progressively fewer electrons are able to reach the anode. When the barrier becomes so high that no one electron can reach it, the negative potential at the anode ceases to rise.

Now we can sum up how the anode current changes in response to changes in the anode voltage in the space-charge-limited region. The decisive factor is the change in the height of the potential barrier near the cathode. When the anode voltage is raised, the barrier is lowered, more electrons are thus able to climb over it, and anode current increases. When the anode voltage is reduced, the potential barrier grows higher, fewer electrons are able to climb over it, more electrons fall back on the cathode, and this means a decrease in the anode current.

The distribution of the space charge in the anode-cathode space can be established from the condition that the anode current density is the same at any section of the electron stream (in the case of plane electrodes) and equal to the product of the space-charge density ρ by the velocity *u*:

$$J_a = \rho u = \text{const} \qquad (12\text{-}3)$$

The equation is quite obvious. The current density increases with increasing number of electrons moving in the unit volume and with increasing electron velocity.

It follows from Eq. (12-3) that

$$\rho = J_a/u \qquad (12\text{-}4)$$

At the anode where electrons have a maximum velocity, the space charge has a minimal density, while the space charge density is a maximum near the cathode where electrons have the lowest velocity.

What we have learned for a diode with plane electrodes fully holds for diodes with any shape of electrodes.

12-2 The Three-Halves Power Law

In a diode operating in the space-charge-limited region the anode current and the anode voltage are connected by a nonlinear relationship which has come to be known as the *three-halves power* (or *Child-Langmuir*) *law*

$$i_a = kv_a^{3/2} \qquad (12\text{-}5)$$

where *k* is known as the perveance of the tube – a constant determined by the geometry of the element structure within the diode tube.

Thus, the anode current increases as the three-halves power of the anode voltage and does not obey Ohm's law. If, for example, we double the anode voltage, the anode current will increase approximately 2.8 times because

$$2^{3/2} = (2^3)^{1/2} \approx 2.8$$

That is, it will be by 40% greater than it should have been in accord with Ohm's law. Graphically, the three-halves power law is depicted by a semi-cubic parabola (Fig. 12-4). Of course, the three-halves power law does not apply to the saturation (or temperature-limited) region at voltages in excess of V_s when $i_a = I_s$ is constant. Sometimes, curve *OAB* in Fig. 12-4 is called the characteristic of the ideal diode.

For a diode with infinite parallel-plane electrodes,

$$k = 2.33 \times 10^{-6}(Q_a/d_{ak}^2) \qquad (12\text{-}6)$$

where Q_a is the effective area of the anode and d_{ak} is the perpendicular distance between anode and cathode.

The actual relation between anode current and anode voltage markedly differs from the three-halves power law. Still, the Child-Langmuir law is convenient to use for analysis because it accounts for the nonlinear behaviour of the tube in a very simple form.

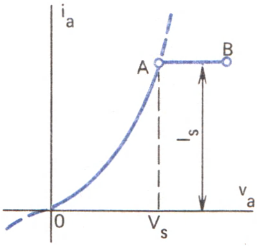


Fig. 12-4

Anode characteristic of an ideal diode or the plot of the three-halves power law

12-3 The Current-Voltage Characteristic of a Vacuum Diode

The performance of a vacuum diode can best be depicted by its *current-voltage characteristic* which relates its anode current to its anode voltage, with the filament (or heater) voltage held constant. The current-voltage characteristic of an actual diode, plotted by experiment (Fig. 12-5), differs from the curve of an ideal diode, which represents the three-halves law and is shown by the dashed curve in the same figure. The difference between the two curves stems from the fact that several assumptions are made in deriving the Child-Langmuir law. Among other things, the initial current I_0 is arbitrarily assumed to be nonexistent and the characteristic is shown starting at the origin (point *O*).

As the filament voltage is raised, point *A* shifts to the left because the emitted electrons acquire a higher initial velocity. The region *BC* of the curve may approximately be regarded as linear. The portion *CD* shows a gradual transition from the space-charge-limited region to the temperature-limited region. At saturation (the *DE* region), a rise in anode voltage entails a rise in anode current. This is due to the Schottky effect and the additional heating of the cathode by anode current.

In the case of a carburized thoria-coated cathode, the Schottky effect is only slightly felt, and the additional heating by anode current is negligible because tungsten has a low resistance and the anode current is small in comparison with the heating current. The characteristic of a diode using such a cathode runs nearly horizontal in the temperature-limited region (Fig. 12-6, curve *1*). Oxide-coated cathodes display a strong Schottky effect, and the additional heating of the cathode by anode current is noticeable because the oxide coating presents a high resistance and the anode current is comparable in magnitude with the heating current. Also, an oxide-coated cathode usually has scattered hot spots because the oxide coating is anything but uniform. Under the temperature-limited conditions, the anode current of a diode using an oxide-coated cathode grows so much (curve *2*) that the transition from the space-charge-limited region to the temperature-limited region cannot usually be discerned from the characteristic (the temperature-limited region becomes noticeable only when an oxide-coated cathode is operated at less than its normal heating voltage).

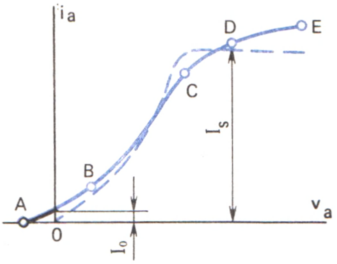


Fig. 12-5

Anode characteristic of an actual diode

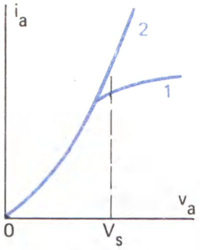


Fig. 12-6

Characteristics of diodes using a carburized thoria-coated cathode and an oxide cathode

12-4 The Parameters of the Vacuum Diode

The parameters of a diode are the quantities that characterize its properties and applicability. Some of them have already been introduced. They are the filament (or heater) voltage V_f, the filament (or heater) current I_f, and the emission current I_e. Now we will define the other parameters.

Slope or a. c. anode conductance, g_a. This quantity, known in US usage as the dynamic plate conductance, g_p, shows how the anode current of a diode changes in response to a change of 1 V in the anode voltage. If the incremental change in anode voltage, Δv_a, brings about an incremental change in anode current equal to Δi_a, the a. c. anode conductance of the diode will be

$$g_a = \Delta i_a / \Delta v_a \qquad (12\text{-}7)$$

The a. c. anode conductance is expressed in milliamperes per volt or in amperes per volt; that is, it has the dimensions of conductance. The term "a. c." or "dynamic" is included because this quantity relates alternating current and voltage. For example, if the a. c. anode conductance is 5 mA V^{-1}, this means that a change of 1 V in anode voltage brings about a change of 5 mA in anode current.

The term “slope” is used because graphically the a.c. anode conductance is the slope of the tangent to the characteristic at a specified point. It is relevant to note that in mathematics the term “slope” defines the ratio of the change in the ordinate to the corresponding change in the abscissa of a point moving along a line. In each particular case, this line may represent some specific behaviour or characteristic and the ratio may result in some specific dimensions, so the slope is usually referred to by some specific name.

The a.c. anode conductance of a diode can be found from its current-voltage characteristic (Fig. 12-7) by taking an incremental change in anode voltage, Δv_a, and the corresponding incremental change in anode current, Δi_a, within some specified region. This procedure is known as construction by *the two-point method.* By definition, the a.c. anode conductance of a diode is the slope of the tangent that makes an angle α with the curve within the chosen region

$$g_a = S_{AB} = k \tan \alpha \qquad (12\text{-}8)$$

Here k is a coefficient which has the dimensions of conductance and takes care of the scale on which the values of current and voltage are laid off.

It would be wrong to write

$$g_a = \tan \alpha$$

because the tangent does not have the dimensions of conductance.

If we recall that the region AB is nonlinear, the slope S_{AB} (or the a.c. anode conductance g_a for that region) has a value which is averaged over the region AB. It is approximately equal to the slope (or a.c. anode conductance) at point Q lying midway between points A and B

$$S_{AB} \approx S_Q$$

It is not convenient to determine the slope or a.c. anode conductance by the two-point method because we have then to draw a tangent, to measure the angle that it makes with the curve, and to account for the scales used in laying off anode voltage and anode current.

On moving into the lower portion of the characteristic (the space-charge-limited region), the a.c. anode conductance falls off, tending to zero. To avoid confusion, it is customary to specify the point or region of the curve for which the quantity is found. For example, one may specify that $g_a = 1.5$ mA V^{-1} at $v_a = 2$ volts.

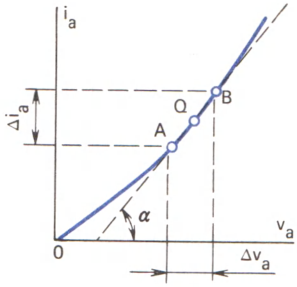


Fig. 12-7

Determining the transconductance of a diode by the two-point method

State-of-the-art diodes have an a.c. anode conductance of 1 to 50 mA V^{-1}. In low-power diodes, it does not exceed several milliamperes per volt. In pulse working, the a.c. anode conductance may be as high as several hundred milliamperes per volt.

The value of g_a depends on the tube design. For a diode with parallel-plane electrodes, it follows from the Child-Langmuir law that

$$g_a = 3.5 \times 10^{-6} (Q_a/d_{ak}^2)\, v_a^{1/2} \qquad (12\text{-}9)$$

Slope or a.c. anode resistance, r_a. This quantity, known in US usage as the dynamic plate resistance, r_p, is the a.c. resistance of the anode-cathode space. It is the reciprocal of the a.c. anode conductance of a diode

$$r_a = \Delta v_a/\Delta i_a = 1/g_a \qquad (12\text{-}10)$$

and is usually as high as hundreds or, sometimes, tens of ohms.

Tubes of higher power ratings have a lower a.c. anode resistance. On moving into the lower region of the V-I characteristic, r_a increases, tending to infinity at the origin of the characteristic.

The a.c. anode resistance of a diode can be found from its V-I characteristic in much the same way as its a.c. anode conductance. This can best be done by the two-point method which gives the average value of r_a for a specified region of the characteristic. If the specified region is only slightly nonlinear, the value of r_a thus found may be taken as its value for the point midway between the extreme points of the region.

The a.c. anode resistance r_a ought not to be confused with the d. c. resistance of a diode, defined as

$$R_0 = v_a/i_a \tag{12-11}$$

and called the *static anode resistance*. As a rule, R_0 is somewhat greater than r_a. From the three-halves law it follows that

$$R_0 = \frac{3}{2} r_a$$

In practice, however, the relation may be different. For example, at $v_a = 0$ and in the presence of some initial anode current, $R_0 = 0$, while r_a at that point will be non-zero. Within the linear region of the diode characteristic, r_a differs only slightly from R_0. The value of r_a decreases with decreasing distance between the anode and cathode and with increasing effective anode area. An idealized dependence of r_a on the electrode geometry may likewise be deduced from the three-halves power law. The resultant expression will then be the inverse of Eq. (12-9).

12-5 Dynamic Operation of a Vacuum Diode, Rectification of Alternating Current

The dynamic operation of a vacuum diode can be analysed grapho-analytically in much the same manner as that of a semiconductor diode (see Chap. 3). With E_a and R_L specified or chosen in advance, we construct the load line, and its intersection with the diode characteristic will define the anode current and the anode voltage. The voltage drop across a vacuum diode may not be usually neglected because it may be as high as tens or even hundreds of volts, depending on the diode design.

Everything we have learned about rectification and simple rectifiers using semiconductor diodes may be applied to rectification and rectifiers using vacuum diodes. A distinction of the latter is the absence of a reverse current. Also, vacuum diodes used as rectifiers at power (commercial) frequency are able to operate at a reverse voltage of several hundred or even several thousand volts. Therefore, there is no need to connect several rectifying diodes in series in order to set up a rectifier capable of withstanding an extremely high reverse voltage.

An undesirable condition for rectifying diodes is a short-circuited load because all of the source voltage would then be applied to the diode, and its anode current would exceed its absolute maximum rating. If this condition is allowed to happen, the cathode will be overheated and damaged or even destroyed completely. The anode will also be overheated. The overheated electrodes give up dissolved or occluded gases, the vacuum in the tube is impaired, the gas is ionized, and the positive ions bombard the cathode, thus aggravating the heat build-up and damage still more.

In rectifying currents at very high frequencies, there is a detrimental effect from the anode-cathode capacitance, C_{ak}, of the diode. It consists of the capacitance between the two electrodes and the capacitance between their leads. In low-power diodes, C_{ak} is several picofarads. At low frequencies, this capacitance does not produce any shunting effect because its reactance is millions of ohms. At frequencies of tens of megahertz and higher, the capacitive reactance becomes comparable with, or even less than, the diode resistance. As a result, alternating current is able to pass through this capacitance, and the rectifying action of the diode is impaired. For example, if $r_a = 500\ \Omega$ and $C_{ak} = 4$ pF, then at a frequency of 200 Hz the capacitive reactance of the diode will be

$$x_C = 1/\omega C_{ak} = 10^{12}/(2\pi \times 200 \times 4)$$
$$\approx 200 \times 10^6\ \Omega = 200\ \mathrm{M\Omega}$$

Practically no current can flow through this reactance. On the contrary, at $f = 200$ MHz, the capacitive reactance x_C is 200 Ω and will shunt the diode heavily.

The value of C_{ak} can be reduced by reducing the size of the anode and cathode and by spacing their leads farther apart. To avoid the likely increase in r_a, the anode-to-cathode spacing is made shorter. Such diodes can handle low values of power.

When using a diode, it is important not to exceed its absolute maximum ratings.

Maximum anode dissipation. If the anode intercepts a number N of electrons per second and each electron has an energy $mu^2/2$, the power imparted by the electron stream and dissipated as heat will be

$$P_a = Nmu^2/2 \tag{12-12}$$

Electrons receive their power from the accelerating field. On neglecting their initial energy, we may write

$$mu^2/2 \approx ev_a$$

Then,

$$P_a = Nev_a \tag{12-13}$$

The product Ne is the quantity of electricity incident on the anode per second, that is, the anode current i_a. Therefore, we finally have

$$P_a = i_a v_a \tag{12-14}$$

The power P_a is the lost power, that is, the power dissipated as heat, and it is technically termed as the *anode dissipation* of a diode. It is not an absolute maximum rating of a diode because it may take on widely varying values depending on the anode voltage. For example, at $v_a = 0$, $P_a = 0$. The anode may also be heated by the thermal radiation emitted by the cathode, but P_a takes into account only the power due to electron bombardment. The greater the value of P_a, the greater the heating of the anode. The anode may be raised to a red heat or even melted.

The value of $P_{a\ max}$ depends on the dimensions, structure and material of the anode and the manner in which it is cooled. It may range from a fraction of a watt for low-power diodes to many kilowatts for high-power tubes. If the anode is not to be overheated, it is essential to satisfy the following condition

$$P_a \leqslant P_{a\ max} \tag{12-15}$$

In pulse working, the instantaneous anode dissipation may exceed $P_{a\ max}$, but the average power ought not to exceed $P_{a\ max}$.

Absolute maximum anode current in pulse working. As a rule, diodes deliver their anode current in the form of pulses. For diodes using oxide-coated cathodes, the limit is set by the destruction of the oxide coating. Each type of diode has an absolute maximum anode current in pulse working of its own, designated as $I_{a\ max}$. Its value is very high for diodes specifically designed for use in pulse circuitry, and it increases with decreasing pulse duration and increasing pulse spacing.

Absolute maximum rectified current. The pulsating current delivered by a rectifying diode has a direct component, $I_{a\ av}$, called the rectified current.

The absolute maximum rectified current, $I_{a\ av\ max}$, is an important rating of diodes. In tubes using an oxide-coated cathode, the emission may be very strong, and for them $I_{a\ av\ max}$ is limited to the value at which the anode current can overheat the oxide coating. In a properly designed diode, this limit corresponds to the absolute maximum anode loading ($P_a = P_{a\ max}$). For diodes used in ordinary rectifiers, it is usual that $I_{a\ av\ max}$ is approximately equal to one-third of $I_{a\ max}$. In operation involving short pulses separated by long intervals, $I_{a\ av\ max}$ is a small fraction of $I_{a\ max}$.

Absolute maximum reverse voltage. When a diode is operating as a rectifier, a negative anode voltage, called the *reverse voltage*, is applied to the diode for some time (a fraction of a cycle). The absolute maximum reverse voltage, $V_{r\ max}$, is a very important parameter. The actual reverse voltage must always be lower than the absolute maximum reverse voltage rating

$$V_r \leqslant V_{r\ max} \tag{12-16}$$

Should V_r exceed $V_{r\ max}$, it is very likely that the leakage currents will increase abruptly, the insulation may break down, field emission from the anode may take place, and the diode may fail.

For low-power rectifying diodes, $V_{r\ max} = 500$ to 1800 V. High-voltage kenotrons have a $V_{r\ max}$ of tens of kilovolts. In them, the anode cap is located at the top of the envelope, that is, as far away from the cathode lead as possible. For low-power diodes, $V_{r\ max}$ is not over 500 V.

At the maker's, diodes are usually tested by causing several devices from each lot to break down when a suitable test voltage is applied. The figure which is one-half to one-third of this test voltage is then taken as $V_{r\ max}$. In this way, if a diode is operated so that its V_r remains lower than or equal to $V_{r\ max}$, an ample reserve of electric strength will be assured, and the tubes will operate reliably.

12-6 Diode Types

Low-power diodes usually have indirectly heated cathodes. In diodes intended for operation at high and very high frequencies, measures are taken to make C_{ak} as small as practicable. Rectifying diodes may have both directly heated and indirectly heated cathodes. Wide use is made of dual diodes (two diode units in the same envelope).

A diode using a directly heated cathode is the simplest of all (Fig. 12-8*a*). This class includes some HV kenotrons and most high-power kenotrons. Indirectly heated cathodes have one of

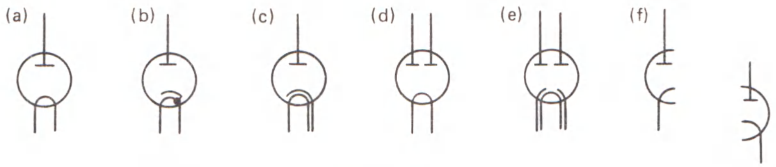


Fig. 12-8
Graphical (circuit) symbols of various diodes

their terminals tied together with one of the heater terminals (Fig. 12-8*b*). Some diodes have a separate cathode lead (Fig. 12-8*c*).

Dual diodes having directly heated cathodes are depicted in a simplified way in Fig. 12-8*d*. Actually they have each two cathodes connected in parallel or in series.

The most commonly used dual diodes with separate cathodes have separate cathode leads (Fig. 12-8*e*). These diodes are often used in two different sections of the same circuit or device. In such cases, it is customary to show a half of a dual diode in the respective section of the circuit (Fig. 12-8*f*). When a dual diode is to be used as a detector, it is fitted with a metal shield to avoid stray capacitive coupling between the two halves. The shield has a lead of its own. Simplified circuit symbols often omit the shield.

Chapter Thirteen
Vacuum Triodes

13-1 Physical Processes

The cathode and anode (or plate) of a triode operate in the same manner as they do in a diode. In the space-charge-limited region, a potential barrier is formed near the cathode. The height of this barrier has a direct bearing on the magnitude of cathode current. The control action of the grid in a triode is similar to that of the anode in a diode. A change in the grid voltage brings about a change in the barrier height. This leads to a related change in the number of electrons capable of overcoming the barrier, that is, in the magnitude of cathode current. If the grid voltage is made more positive, the potential barrier is reduced, a greater number of electrons are now able to clear it, and the cathode current rises. When the grid becomes negative, the potential barrier gains in height, fewer electrons are able to climb over, and the cathode current falls.

Current control by the grid in a triode is similar to current control in a bipolar transistor. In a transistor, a change in the voltage across the emitter junction brings about a change in the height of the respective potential barrier. As a result, the emitter current is changed. However, the grid not only controls the cathode current, but also substantially affects the behaviour of the anode. For the electric field set up by the anode the grid acts as an electrostatic shield, that is, as an obstacle (provided the grid is connected to the cathode). The greater proportion of the flux due to the anode is intercepted by the grid, and only a negligible part of the flux penetrates the grid and reaches the potential barrier near the cathode. In this way, the grid shadows the cathode from the anode and minimizes the effect that the anode might have had on the potential barrier at the cathode. The grid is said to intercept the greater proportion of the electric flux produced by the anode.

The shadowing effect of the grid is visualized in the diagram of Fig. 13-1*a* which shows the electric field pattern for a triode with parallel-plane electrodes and with the grid short-circuited to the cathode, that is, when $v_g = 0$. For simplicity, the space charge is neglected. As is seen, the grid intercepts the greater proportion of the flux emerging from the positive anode. In other words, the grid weakens the effect of the

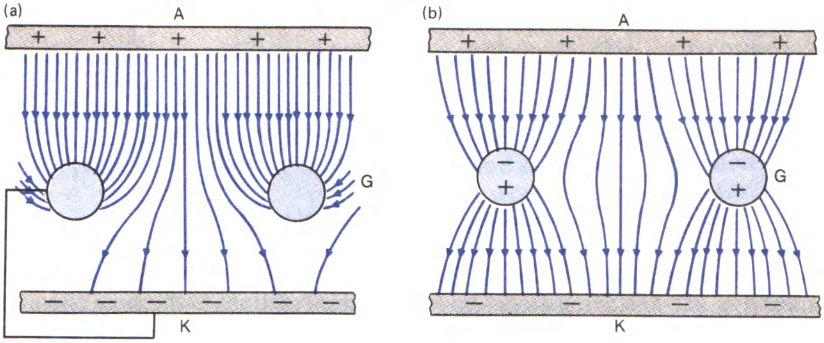


Fig. 13-1
Electric field in a triode: (*a*) at zero grid voltage and (*b*) with an isolated grid

anode on the cathode in a triode. If, however, the grid is not connected to the cathode and isolated from the other electrodes, it will not be able to weaken the field at the cathode. Then (Fig. 13-1*b*) two equal unlike image charges are produced at the grid due to electrostatic induction, and the field at the cathode has the same intensity as it would without a grid.

Most of the flux fails to reach the surface of the cathode, but terminates at the electrons that form the space charge cloud near the cathode. For simplicity, here and elsewhere we will speak of the penetration of the flux towards the cathode, meaning that actually the flux acts upon the electrons of the space charge. The denser the grid (that is, the shorter the pitch of grid wires), or the greater the wire diameter (that is, the smaller the spacing between the grid wires), the smaller the fraction of the anode flux that is able to penetrate the grid. The shadowing effect of the grid is a maximum at some optimal location of the grid between the anode and cathode.

To sum up, the grid weakens the effect of the anode more as its pitch is reduced. In diodes, normal values of anode current are obtained at anode voltages of several volts or 20 to 30 volts. In the presence of a grid and at $v_g = 0$, similar magnitudes of anode current are obtained at anode voltages of tens or even hundreds of volts.

The grid affects the anode current far more strongly than does the anode. If we apply a voltage to the grid, the electric field set up by the grid will be free to reach the cathode because there is no obstacle for the field between the grid and the cathode. The grid holds a 'dominant' position. It exerts a strong action on electrons while that of the anode is reduced to a small fraction because only a small proportion of the total anode flux is able to penetrate the grid. It would be wrong to assert that the grid produces a stronger action than the anode for the reason that it is located closer to the cathode. If we place the grid near the anode so that it is only slightly closer to the cathode than the anode, it will weaken the anode flux reaching the cathode many times all the same. Thus, the proximity of the grid to the anode is not the principal factor that controls the anode current.

How many times more effective the grid is than the anode in controlling the anode current is stated in terms of what is known as the *amplification factor* of a triode, symbolized as μ. If $\mu = 10$, this means that the grid controls the anode current ten times stronger than the anode. The smaller the grid pitch, the greater the value of μ. With a grid of a given pitch, the amplification factor μ is a maximum at some optimal location of the grid between anode and cathode. The μ of state-of-the-art triodes is from several units to several tens.

Some authors use what is called the *penetration factor* which is symbolized as *D* (from the German 'Durchgriff' for penetration). For triodes, it is the reciprocal of the amplification factor

$$D = 1/\mu \qquad (13\text{-}1)$$

Obviously, *D* is less than unity. The penetration factor shows what fraction of the grid's action is equivalent to the anode's action on the cathode current. If, for example, $\mu = 10$, then $D = 0.1$. This means that the anode's action is 0.1 of the grid's action.

The term 'penetration factor' (Durchgriff) was first introduced by Heinrich Barkhausen of Germany who made a valuable contribution to the theory of vacuum tubes. It stresses the shadowing effect of the grid. It may be said that it characterizes the grid's transparency for the electric field set up by the anode. The wider the pitch of the grid wires, the easier it is for the flux lines to pass from the anode through the grid to the cathode, and the greater the value of D, but the amplification factor μ decreases in proportion. It would be wrong to treat D as characterizing the grid's transparency for the electron stream. Of course, a denser grid presents a greater obstacle to the electron stream, but this does not mean that D shows what fraction of the stream passes through the grid.

Of special interest is what happens in a vacuum triode when the grid is negative because receiving-amplifying tubes are usually operated in this condition. The negative grid produces in the grid-cathode space a retarding field which opposes the accelerating field that extends from the anode. As a result, the potential barrier at the cathode grows higher, and the cathode current is decreased, At some negative grid voltage, the current falls to zero, and the tube is said to be driven to *cutoff*. Accordingly, the respective negative grid voltage is termed the *cutoff bias voltage*, $v_{g\ co}$. At this voltage, the field set up by the grid in the grid-cathode space raises the potential barrier so much that all electrons emitted by the cathode fall back. If, however, at $v_g < 0$, the tube is not driven to cutoff, then this means that electrons having a high initial velocity do overcome the potential barrier and travel towards the anode.

The grid cutoff bias voltage is small in comparison with the anode voltage because the grid produces a stronger controlling action than does the anode. For example, in a triode with $\mu = 20$, the cutoff bias voltage will be -5 V at $v_a = 100$ V. At $\mu = 20$, an anode voltage of 100 V is equivalent in its action to a grid voltage of 5 V. Therefore, the effect of the anode can fully be cancelled by applying $v_{g\ co} = -5$ V to the grid.

To sum up, a relatively small negative voltage applied to the grid is able to cut down or even completely stop the flow of anode current.

A positive voltage applied to the grid will set up an accelerating field which is added to the field extending from the anode. The resultant field brings down the potential barrier, more electrons are now able to climb over it, and the cathode current rises. Some of the electrons will however be inevitably attracted by the positive grid, and a current will be flowing in the grid circuit. This *grid current* is undesirable and even harmful in many cases. If the positive voltage at the grid is a small fraction of the anode voltage, the grid current will be negligibly small in many cases. The denser the grid and the more positive it is, the higher the grid current.

Since the grid acts far more strongly than does the anode, even a relatively small positive voltage applied to the grid can bring about a substantial rise in the anode current. As an example, let a triode have $\mu = 20$, $v_g = 0$, $v_a = 100$ V, and an anode current of 10 mA. Then for the anode current to double (to rise to 20 mA), while the grid voltage is held constant, the anode voltage should likewise be doubled, that is, raised to 200 V. However, at $\mu = 20$, an anode voltage of 100 V is equivalent in terms of the controlling action to a grid voltage of 5 V. Therefore, we may achieve our goal by applying $+5$ V to the grid rather than double the anode voltage – the anode current will then rise to the desired 20 mA.

To sum up, making the grid more positive leads to a rise in both the anode and grid currents.

When the grid is made very positive, the grid current may rise so much that the anode current will sometimes fall.

By varying the grid voltage from its cutoff negative value, we can control the anode current between broad limits – from zero to some maximum value. Such is the control action of the grid. Importantly, substantial changes in anode current can be brought about by relatively small changes in grid voltage. If we wished to achieve a comparable change in anode current, the changes in anode voltage would have to be μ times greater. In other words, small changes in grid voltage are equivalent to μ times changes in anode voltage. This basic property of vacuum triodes makes them ideally suited for amplification.

The performance of vacuum triodes is markedly affected by what is known as the *island effect*. It consists in that because the grid structure is nonuniform, the field set up by the grid is likewise nonuniform. Therefore, the height of the potential barrier varies from one point to another at the cathode. When the grid voltage falls to a certain minimum value, the emission from the cathode is restricted to a few small areas, or islands, of the cathode. This effect gains in strength as the tube approaches cutoff. The effect is also stronger with a grid made of wires spaced wider apart and located closer to the cathode.

13-2 Current Division

When the grid is positive, the division of the cathode current takes place in a vacuum triode between the grid and the anode. If the anode voltage exceeds the grid voltage, some electrons will be intercepted by the grid, and the remainder will pass through the grid to the anode. The triode is said to be operating in the *intercept* (or *current division*) *mode*. In this mode, the grid current is a small fraction of the anode current. If, on the other hand, the grid voltage is about equal to, or greater than, the anode voltage, many of the electrons passing through the grid will be retarded in the grid-anode space, their paths will be strongly curved, their axial component of velocity will reduce to zero, and the electrons will go back to the grid. The tube is said to be operating in the *fall-back mode*. Obviously, the fall-back is always accompanied by the intercept mode.

Figure 13-2 shows most typical paths travelled by electrons in the fall-back mode. Electrons *1*, *2* and *3* are intercepted by the grid. In fact, electron *3* has its path bent by the grid so that it cannot pass through the grid, and so it impinges on the grid. Electrons *5* and *6* pass through the grid and reach the anode while electron *4* falls back to the grid. In going back to the grid, electron *7* moves past the grid wires, enters the grid-cathode space, is retarded there, moves again towards the grid to finally fall on it.

At $v_a = 0$ and $v_g > 0$, a second potential barrier is formed between the grid and anode. It is labelled as *II* in the figure (barrier *I* exists near the cathode). In the circumstances, nearly all the electrons passing through the grid go back because they are not able to climb over the second potential barrier. This is the reason why at $v_a = 0$ the grid current has a maximum value. Only very few electrons are able to climb over the second potential barrier and reach the anode, thus producing the initial anode current.

If we apply a positive voltage to the anode now, the second potential barrier will be reduced in height, more electrons will be able to climb over it, and the anode current will rise. The electron space charge in the second potential barrier and the anode make up a system similar to a diode. The effect of the anode field on this electron space charge is not weakened any longer, and the anode current abruptly rises even at small positive anode voltages while the grid current abruptly falls because fewer electrons go back to the grid. A sudden change takes place in the division of the cathode current between the grid and anode – a situation typical of operation in the fall-back mode.

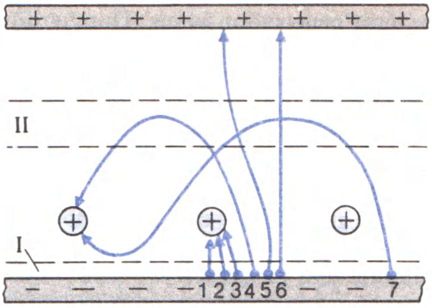


Fig. 13-2

Travel path of electrons in the fall-back mode

At some positive anode voltage, the second potential barrier is reduced in height so much that none of the electrons go back to the grid. The tube has moved into the intercept mode. The further rise in anode voltage brings about a rise in anode current because the anode field lowers the potential barrier at the cathode and also because there is a change in current division. Now, however, the anode current builds up at a slower rate because the effect of the anode field on the potential barrier at the cathode is weakened by the grid. The grid current decreases likewise only slightly because the number of electrons moving from the cathode directly to the grid wires only slightly depends on the anode voltage.

The ratio v_a/v_g at which a transition from the intercept and fall-back modes takes place is different for different tubes and depends on the electrode structure.

Current division is stated in terms of the *current division ratio* defined as

$$k_d = i_a/i_k = i_a/(i_a + i_g) \tag{13-2}$$

which can never be greater than unity and which shows what fraction of the cathode current is the anode current.

The current division ratio is a function of the ratio v_a/v_g and the grid design. For example, a denser grid will have a lower k_d because a denser grid will intercept more electrons. The manner in which k_d varies with v_a/v_g is shown in the plot of Fig. 13-3. At $v_a = 0$, $v_a/v_g = 0$, and k_d has a minimal value close to zero because there is only a small anode current flowing due to the initial velocity of electrons. As v_a/v_g is increased, k_d rapidly increases at first, which corresponds to the fall-back mode (region *I*). As the tube moves into the intercept mode (region *II*), it grows at a lower rate, tending to unity.

13-3 The Virtual Voltage and the Three-Halves Power Law

The cathode current of a vacuum triode can be calculated on replacing the tube with an equivalent diode. The procedure is as follows. If, in a triode, we place what is known as a virtual anode instead of the grid, the anode current in such a diode will, at some anode voltage, be equal to the cathode current of the original triode (Fig. 13-4). This is the *equivalent*, *virtual* or *composite anode voltage*, $v_{a\ eq}$, defined as

$$v_{a\ eq} = v_g + Dv_a = v_g + v_a/\mu \qquad (13\text{-}3)$$

The above equation may be interpreted as follows. The field set up by the grid is acting at full strength, and the field set up by the anode in the grid-cathode space is attenuated by the shadowing effect of the grid. This shadowing effect is expressed in terms of *D*, the penetration factor, which is the reciprocal of the amplification factor μ.Therefore, v_g may not be added directly to v_a; at first v_a must be multiplied by *D* or divided by μ. Equation (13-3) is an approximate one because it does not take into account the island effect. If the grid wires are spaced not very far apart, the error will be small. For practical purpose, Eq. (13-3) is accurate enough.

In the equivalent diode, the anode current is equal to the cathode current of the original triode, and the actual anode voltage is replaced by the equivalent anode voltage. Therefore, the three-halves power law for a vacuum triode may be re-written thus:

$$i_k = kv_{a\ eq}^{3/2} = k\,(v_g + Dv_a)^{3/2} \qquad (13\text{-}4)$$

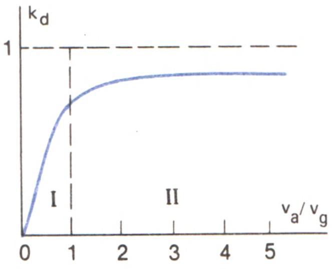


Fig. 13-3

Current division ratio as a function of v_a/v_g

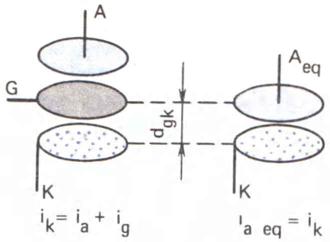


Fig. 13-4

Reducing a triode to an equivalent diode

Recalling that in an equivalent diode the virtual anode takes the place of the grid in an actual triode, we obtain for a triode with parallel-plane electrodes

$$k = 2.33 \times 10^{-6} Q_a/d_{gk}^2 \qquad (13\text{-}5)$$

The surface area of the virtual anode in an equivalent diode is equal to the surface area of the actual anode. In an implicit form, Eq. (13-4) also includes the anode-cathode spacing and the dimensions defining the grid density because these quantities control the penetration factor.

The three-halves power law for triodes is a very approximate one. Still, it is useful when the performance of a triode is examined theoretically. For practical purposes, resort is made to characteristics measured by experiment and published in reference sources.

Invoking the three-halves power law, we can find the grid cutoff bias voltage, $v_{g\ co}$, at a given anode voltage, v_a. If the tube is at cutoff, then $i_k = 0$. It follows from the three-halves power law that this can happen only if

$$v_{a\ eq} = v_{g\ co} + Dv_a = 0 \qquad (13\text{-}6)$$

On solving Eq. (13-6) for $v_{g\ co}$, we get

$$v_{g\ co} = -\,Dv_a \text{ or } v_{g\ co} = -\,v_a/\mu \qquad (13\text{-}7)$$

The actual cutoff bias voltage is somewhat higher than the value given by Eq. (13-7). This is explained mainly by the fact that the island effect causes D to rise somewhat on approaching cutoff.

13-4 Characteristics of the Vacuum Triode

A characteristic is a relationship between two quantities which characterize the behaviour of a device, circuit, or equipment. They are usually plotted in the form of families of graphs (characteristic curves) relating the currents obtained to the voltages applied for a range of operating conditions.

The electrode characteristic shows the relationship between current and voltage at an electrode of the device, say, anode current against anode voltage in a tube. The transfer characteristic shows the relationship between the current (or voltage) at one electrode and the voltage (or current) at some other electrode, say anode current and grid voltage. The static characteristic shows, for example, anode current against grid voltage, with all other applied voltages held constant and with no load connected, that is, under static conditions. The dynamic characteristic relates the current from one electrode and the voltage at another under dynamic conditions.

The characteristics of a triode may be plotted on the basis of the three-halves power law – they will then be theoretical, or idealized, characteristics. Since they are plotted under a number of simplifying assumptions, they are not accurate. Actual characteristics are measured by experiment. They are more accurate because they take into account the island effect, variations in temperature at various points on the cathode, the nonequipotential surface of directly heated cathodes, the Schottky effect, the initial velocity of electrons, the additional heating of the cathode by the anode current, the contact potential difference, the thermo-emf generated on heating the junction of two dissimilar metals, and other phenomena none of which are reflected in the three-halves power law.

The characteristics given in reference sources (data sheets or handbooks) are the average ones – that is, deduced from those measured for a number of tubes of the same type. Therefore, their use leads to further errors.

We will begin our study of characteristics by considering the static characteristics of the vacuum triode.

The anode current of a triode is a function of grid and anode voltages

$$i_a = f(v_g, v_a) \qquad (13\text{-}8)$$

The same is true of the grid and cathode currents of a vacuum triode

$$i_g = f_1(v_g, v_a)$$

and (13-9)

$$i_k = f_2(v_g, v_a)$$

Any relationship between three quantities can only be shown graphically in a three-dimensional (or spatial) coordinate system, which is an obvious inconvenience. Common practice is to plot two-dimensional graphs, one of the voltages being held constant and considering the current as a function of only the other voltage.

It is widely practiced to relate the anode, grid or cathode current of a triode to its grid voltage, with the anode voltage, v_a, held constant

$$\begin{aligned} i_a &= F(v_g) \\ i_g &= F_1(v_g) \qquad (13\text{-}10) \\ i_k &= F_2(v_g) \end{aligned}$$

The first two relationships are most important. The curves depicting the relationship between anode current and grid voltage, $i_a = F(v_g)$, are called the *static anode current/grid voltage characteristics* of a triode. They are similar to the static transfer characteristics of a transistor. The curves depicting the relationship between the grid current and the grid voltage of a vacuum triode, $i_g = F_1(v_g)$, are called the *static grid characteristics*. In the case of transistors, they are called the *static input characteristics*. There is a separate curve for each value of anode voltage, so for each value of current there is a family of characteristics. They are plotted for anode voltages spaced a certain interval apart.

Alternatively, we may plot a curve relating the anode, grid or cathode current to the anode voltage of a vacuum triode, with its grid voltage, v_g, held constant

$$\begin{aligned} i_a &= f(v_a) \\ i_g &= f_1(v_a) \qquad (13\text{-}11) \\ i_k &= f_2(v_a) \end{aligned}$$

Here, the most important characteristics are the *anode characteristics* similar to the output cha-

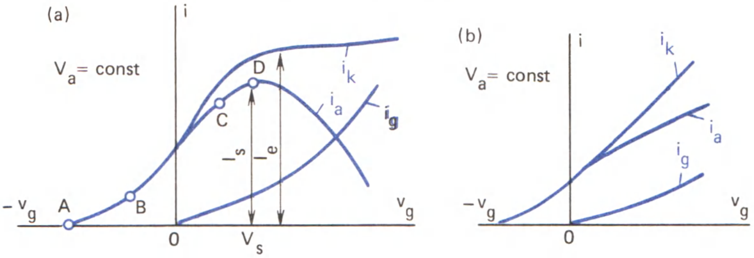


Fig. 13-5

Anode, grid and cathode current characteristics of a triode

racteristics of transistors and relating the anode current to the anode voltage of a vacuum triode, $i_a = f(v_a)$, and the *grid-anode* (or *inverse transfer*) *characteristics* relating the grid current to the anode voltage of a vacuum triode, $i_g = = f_1 (v_a)$.

As a rule, tube data sheets and reference books only give families of characteristics for anode and grid currents. Characteristics for cathode current can readily be derived by simply adding the ordinates of the first two characteristics.

Where the anode current of a vacuum triode is to be calculated for practical purposes, it will suffice to have only a family of either anode-grid or anode characteristics. The former visualize more readily the control action of the grid, so some authors call them the *control characteristics* of a tube. Calculations based on anode characteristics are simpler to do and they yield more accurate results.

Figure 13-5*a* shows characteristics relating the anode, grid and cathode currents of a triode to its grid voltage with its anode voltage held constant. They obviously hold for the operation of the tube (with, say, a tungsten cathode) in the temperature-limited or saturation region. At $v_g < 0$, the characteristics for anode and cathode current run together. Owing to the island effect and other factors, the start of a curve (point *A*) usually represents a lower cutoff bias voltage than that found by Eq. (13-7).

As we reduce the grid voltage in absolute value, the tube is finally driven into conduction (it is said to be cut in), the potential barrier at the cathode becomes lower in height, and there is a rise in the anode current of the tube. The number of electrons able to climb over the potential barrier at the cathode increases in a nonlinear fashion, and so the characteristic has a nonlinear tail, *AB*, which gradually blends into the middle, approximately linear portion, *BC*. When the grid is at a positive potential, the characteristic for cathode current runs above that for anode current because the grid draws some current. The characteristic for grid current starts at the origin of coordinates, similarly to the diode characteristic.

As the grid is made progressively more positive, all currents rise at first. The upper portion of the characteristic for anode current, *CD*, corresponds to a gradual transition into the temperature-limited region. At saturation, any further rise in the grid voltage brings about only a slight increase in the cathode current, but the grid current builds up, and this causes a fall in the anode current. When the grid is at a very high positive potential, the anode current becomes smaller than the grid current.

In tubes using activated (say, oxide-coated) cathodes, the cathode current in the temperature-limited region (at saturation) builds up in about the same manner as it does in the space-charge limited region. If now the grid current is raised at a slower rate than the cathode current, the characteristic for the anode current will have an up-sloping portion (Fig. 13-5*b*). If, on the other hand, the grid current is raised faster than the cathode current, the anode current will fall. The denser the grid and the smaller the anode voltage, the higher the rate at which the grid current rises.

Only transmitting and switching tubes are operated with a very high positive voltage at their grids. In receiving-amplifying tubes, the grid is usually held negative all the time. Reference sources often give the characteristics of receiving-amplifying tubes for only a negative voltage at the grid.

The anode-grid characteristic of a tube may take up a position which depends on how dense the grid is or, which is the same, on how high its amplification factor μ is. With a dense grid (a high value of μ), the grid cutoff bias voltage of the tube is low, and the most of the characteristic lies within the region of positive grid voltages. Sometimes, such curves are referred to as high-cutoff characteristics and the tubes themselves as high-mu tubes. With an open grid (that is, one with the grid wires spaced wide apart and with a low value of μ), the grid cutoff bias voltage is high and the characteristic lies in the region of negative voltages. Sometimes, such curves are referred to as low-cutoff characteristics and the tubes themselves as low-mu tubes. Low-mu tubes are able to handle a considerable anode current and draw no grid current.

Families of anode-grid and grid characteristics of a vacuum triode are shown in Fig. 13-6. When the anode voltage is raised, the anode-current curve is shifted to the left and the grid-current curve runs lower. This can be explained as follows. The higher the anode voltage, the higher (in absolute value) the grid cutoff bias voltage, and the higher the anode current at a given grid voltage. The grid current, however, decreases because the stronger anode field prevents many electrons from being attracted by the grid. Conversely, when the anode voltage is reduced, the grid attracts more electrons, and its current builds up. The grid current curve takes up the topmost position at $v_a = 0$.

Sometimes we may need additional curves not included in the family of characteristics (they are shown dashed in the figure), for example, the curve at an anode voltage of $(V_{a2} + V_{a3})/2$. A curve lying outside the available family is plotted on the assumption that, approximately, it is shifted in proportion to the anode voltage. As an example, the figure shows the curve plotted at V_{a4} such that

$$V_{a4} - V_{a3} = V_{a3} - V_{a2} = V_{a2} - V_{a1}$$

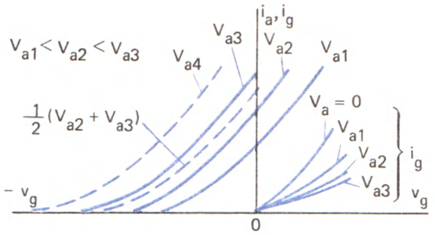


Fig. 13-6

Families of anode-grid and grid characteristics of a triode

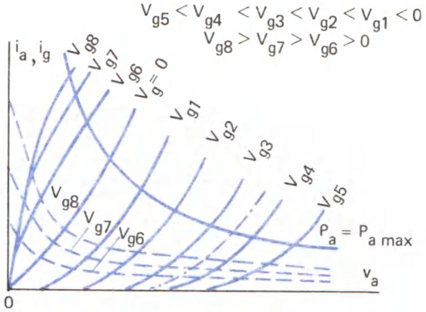


Fig. 13-7

Families of anode and grid-anode characteristics and the maximum anode dissipation curve

Let us take a closer look at families of anode and grid-anode characteristics (Fig. 13-7). The anode characteristic at $V_g = 0$ starts at the origin of coordinates. At lower values of grid voltage, V_{g1} through V_{g5}, the anode characteristics are shifted to the right (because a higher cut-in voltage is needed) and diverge a little. In contrast to idealized curves, the actual characteristics are not shifted in precise proportion to the grid voltage. The anode characteristics plotted at positive grid voltages V_{g6}, V_{g7} and V_{g8} start at the origin of coordinates to the left of the curve for $V_g = 0$ and bulge to the left rather than to the right. At first, they steeply slope upwards, then they level off, and the slope of the curves decreases.

The grid-anode characteristics (shown dashed in the figure) are given only for positive voltages at the grid because a negative grid does not draw any current. At $v_a = 0$, the grid current is a maximum and its value increases in proportion to the grid voltage. As the anode voltage is raised, the grid current abruptly falls at first (in the fall-back mode) because of the current division, following which (in the intercept mode) it falls off slightly.

The family of anode characteristics may often include a line representing the absolute maximum anode dissipation. Since $P_a = i_a v_a$, the line may be described by an equation of the form

$$i_a = P_{a\ max}/v_a \tag{13-12}$$

If we know $P_{a\ max}$ and have several values of anode voltage, we can readily calculate the

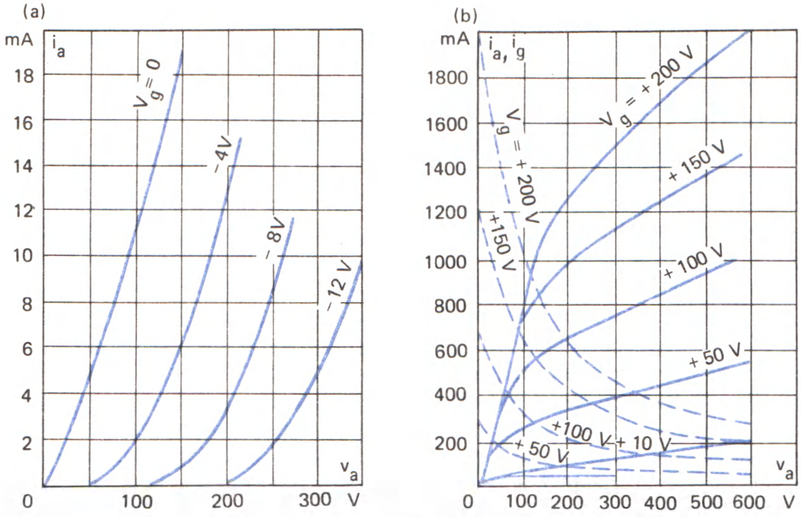


Fig. 13-8

Family of the anode characteristics of a triode at negative grid voltages (*a*) and its pulse characteristics at high positive grid voltages (*b*)

respective anode current and draw a $P_{\text{a max}}$ curve which will be a hyperbola. The area above this curve is the region where the tube cannot be operated at continuous currents such that $P_a >$ $> P_{\text{a max}}$. In pulse working, the operation within the region above the $P_{\text{a max}}$ curve is permissible, provided the average power dissipated at the anode does not exceed its absolute maximum rating.

Additional curves may likewise be added to a family of anode characteristics. As an example, the figure shows a dashed curve for a voltage lying midway between V_{g3} and V_{g4}.

In pulse working, the anode currents may be many times their values in CW working. This condition is obtained by feeding short high-value pulses of voltage to the anode and grid. In the case of pulse working, the anode characteristics are plotted for a specified pulse duration τ_p and a specified pulse repetition frequency *f*. An increase in τ_p and *f* leads to a decrease in the anode and grid currents due to the 'poisoning' of the cathode.

Figure 13-8 shows the characteristics of a vacuum triode as plotted for a set of operating conditions. The pulse characteristics (Fig. 13-8*b*) have been plotted for $\tau_p = 2$ μs and $f = 1$ kHz. The small shaded area in the same figure corresponds to the family of curves shown in Fig. 13-8*a*.

13-5 The Operating Conditions, Absolute Maximum Ratings and Parameters of the Vacuum Triode

The *operating conditions* of the vacuum triode include the filament (or heater) voltage V_f, the filament (or heater) current I_f, the normal direct anode and grid voltages and the respective direct anode current. These voltages are not mandatory, and tubes may often operate at some other voltages. For example, a lower anode voltage may well be used. A higher anode voltage may also be used, provided this does not lead to an excessive heat dissipation at the anode.

The *absolute maximum ratings* of tubes are very important quantities because they set limits of usability of a tube. These are the *absolute maximum anode dissipation* $P_{\text{a max}}$, the *absolute maximum grid dissipation* $P_{\text{g max}}$, the *absolute maximum anode voltage* $V_{\text{a max}}$, the *absolute maximum heater-to-cathode voltage* V_{hk}, and the *absolute maximum cathode current* $I_{\text{k max}}$. For triodes intended for use in pulse circuitry, data sheets specify the *absolute maximum pulse anode and cathode currents*.

The analysis of vacuum-tube operation calls for knowledge of certain dynamic factors called *tube parameters* or *constants*. These parameters tell how two of the three important quantities – anode voltage, grid voltage, and anode current – are related while the third is held constant. They have specific names and their values may be determined from the anode characteristics of a tube.

These parameters are the *transconductance* (also known as the *mutual conductance*), the *slope* or *a. c. anode resistance* (the dynamic plate resistance in US usage), and the *amplification factor* or the *penetration factor*. They describe the performance of a tube under static conditions, that is with no load connected to its plate circuit.

The *transconductance*, g_m, of a vacuum triode shows how effective its grid is in controlling the anode current. If, with the anode voltage held constant, a change in grid voltage, Δv_g, brings about a change of Δi_a in anode current, then

$$g_m = \Delta i_a / \Delta v_g \text{ with } v_a \text{ held constant} \quad (13\text{-}13)$$

Or, in words, the transconductance of a triode is the ratio of a slight change in anode current to the slight change in grid voltage that caused it, with the anode voltage held constant while the changes take place. The condition $v_a = \text{const}$ is essential so that the transconductance could reflect the action of only the grid voltage of the tube. The 'trans' part of the word 'transconductance' denotes a transition from one side (the input or grid side) to the other (the output or anode) side of the tube. The alternative term 'mutual conductance' refers to the fact that this quantity is mutual to both the input (grid) and the output (anode) circuits of a tube. The 'conductance' part of the term is due to the fact that current divided by voltage yields the dimensions of conductance.

The transconductance of a vacuum triode is analogous to the y_{21} parameter of bipolar transistors or the transadmittance of FETs.

The transconductance of a tube is expressed in milliamperes per volt or in amperes per volt. It may be said that the transconductance shows how much, in milliamperes, the anode current changes when the grid voltage is changed by 1 volt, with the anode voltage held constant. For example, if $\Delta v_g = 2$ V and $\Delta i_a = 6$ mA, then $g_m = 6 \div 2 = 3$ mA V^{-1}.

There is a difference in the concept of transconductance as applied to a vacuum triode and the concept of a. c. anode conductance as applied to a vacuum diode. In the case of a diode, it is the reciprocal of the dynamic a. c. resistance of the cathode-to-plate space. In the case of a vacuum triode, this quantity is not the reciprocal of the a. c. resistance of the grid-cathode space although it has the dimensions of conductance.

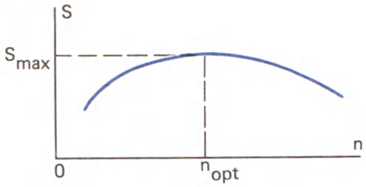


Fig. 13-9
Transconductance as a function of grid-wire pitch

State-of-the-art vacuum triodes have a transconductance of 1 to 50 mA V^{-1}. The greater the transconductance, the better the tube because its grid exerts a stronger controlling action on the anode current. In most cases, the transconductance is several milliamperes per volt; if it is in excess of 10 mA V^{-1}, the tube is said to have a high transconductance.

The value of transconductance depends on the electrode geometry and structure and the operating conditions of the tube. For a triode with parallel-plane electrodes operating at $v_g < 0$, it follows from the three-halves power law that

$$g_m = 3.5 \times 10^{-6} (Q_a / d_{gk}^2)(v_g + Dv_a)^{1/2} \quad (13\text{-}14)$$

As is seen, the transconductance increases with increasing grid and anode voltages and also with increasing surface area of the anode and decreasing grid-to-cathode spacing. The latter factor exerts an especially strong influence. The smaller the value of d_{gk}, the stronger the controlling action of the grid on the height of the potential barrier at the cathode.

When the grid wires are spaced wider apart (this is an 'open' grid), D increases and it would seem, in accord with Eq. (13-14), that g_m should also increase. Actually, an 'open' grid has a reduced effect on the potential barrier owing to the island effect, and g_m is reduced. For each value of d_{gk} there is an optimal grid wire pitch at which the transconductance is a maximum. This is borne out by the plot relating g_m to the number n of grid wires per centimetre of its length (Fig. 13-9).

Now we will find the transconductance of a vacuum triode from its characteristics. The transconductance is related to the slope of the anode-grid characteristic. The greater the slope, the greater the transconductance. In turn, the

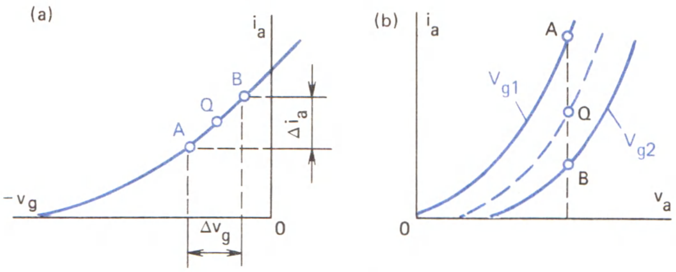


Fig. 13-10
Finding transconductance from characteristics

slope of a curve is defined as the slope of a straight line drawn tangent to the curve at a specified point. The slope of the tangent line is the ratio of the altitude to the base of a right triangle that has the tangent line for its hypotenuse. The simplest way to find the transconductance of a tube is by the two-point method (Fig. 13-10*a*). If the region between points *A* and *B* is nonlinear, the slope of the tangent line found for this region, $S_{AB} = g_{m\ AB}$, is the average over that region and equal to the slope (transconductance) at the middle point *Q* (the quiescent point). Within the tail of the curve the slope (transconductance) decreases and falls nearly to zero at the start of the curve.

The two-point method may also be used in order to find the transconductance of a tube from its anode characteristics (Fig. 13-10*b*). To this end, choose points *A* and *B* for V_{g1} and V_{g2}, corresponding to the same anode voltage. Divide the change Δi_a that takes place in passing from point *A* to point *B* by the respective change in grid voltage, $\Delta v_g = V_{g1} - V_{g2}$. The slope (transconductance) thus found, $S_{AB} = g_{m\ AB}$, is the average for region *AB* and it may be assigned to point *Q* on the middle curve corresponding to $(V_{g1} + V_{g2})/2$. Thus, in order to find the transconductance at point *Q* from the anode characteristics of a tube, use the changes in current and voltage between points *A* and *B* lying on adjacent curves.

The *a.c. anode* (or *dynamic plate*) *resistance*, r_a, shows how anode voltage affects anode current. This tube constant has the same physical meaning as the a.c. anode (or dynamic plate) resistance of a vacuum diode – it is the resistance of the anode-to-cathode space to alternating anode current. When finding r_a for a vacuum triode, however, it is important to hold its grid voltage constant.

If a change of Δv_a in anode voltage causes the anode current to change by Δi_a, the a.c. anode resistance will be

$$r_a = \Delta v_a / \Delta i_a \text{ with } v_g \text{ held constant} \qquad (13\text{-}15)$$

For example, if $\Delta v_a = 50$ V and $\Delta i_a = 2$ mA, the a.c. anode resistance will be $r_a = 50 \div 2 =$ $= 25$ kΩ.

As is seen, the a.c. anode resistance of a triode is the rate of change of anode voltage with respect to anode current, with the grid voltage held constant. The condition $v_g = \text{const}$ is essential for the a.c. anode resistance to reflect the influence of only the anode voltage.

The greater the a.c. anode resistance, the weaker the controlling action of the anode on anode current. In other words, if we wish to obtain the same change in anode current Δi_a at a higher value of r_a, we need to change the anode voltage by a greater amount.

Instead of the a.c. anode resistance, we might well use its reciprocal, called the a.c. anode conductance, g_a:

$$g_a = 1/r_a = \Delta i_a / \Delta v_a \text{ with } v_g \text{ held constant} \qquad (13\text{-}16)$$

In terms of physical significance, it is similar to the transconductance, g_m, of a triode because it shows how much the anode current changes when the anode voltage is changed by 1 volt. Although g_a and r_a are equals, preference in practical usage is given to r_a.

The g_a or $1/r_a$ constant is analogous to the y_{22} parameter of bipolar transistors or the $1/r_{ds}$ $(1/r_d = g_D)$ of FETs.

For vacuum triodes, r_a ranges between 0.5 and 100 kΩ, the most common figures being from several kilohms to 10-30 kilohms. These values of r_a hold for operation of a triode within the linear portions of the characteristics.

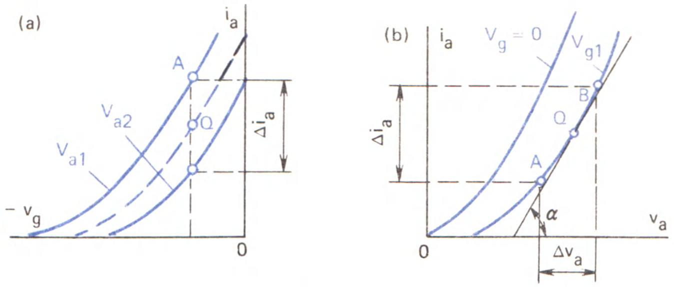


Fig. 13-11
Finding anode slope (a. c.) resistance from characteristics

On the basis of the three-halves power law, we may write the following formula for r_a:

$$r_a = \frac{d_{gk}^2}{3.5 \times 10^{-6}\, Q_a D\, (v_g + Dv_a)^{1/2}} \qquad (13\text{-}17)$$

It is seen that r_a decreases with decreasing d_{gk} and increasing Q_a. If D increases, say, when the grid wires are spaced wider apart, r_a decreases because the anode is in a position to affect the potential barrier and, in consequence, the anode current more strongly through such a grid. The anode-to-cathode spacing, d_{ak}, does not enter Eq. (13-17) explicitly, but an increase in d_{ak} leads to a reduced action of the anode. As a result, r_a rises and D falls. A decrease in D leads to a rise in r_a. The grid wire pitch affects r_a most of all. The shorter the grid wire pitch, the greater the value of r_a. A decrease in grid and anode voltages leads to an increase in r_a because the potential barrier grows higher.

To find r_a from the anode-grid characteristics of a tube, hold the grid voltage constant and measure the increase in the anode current, Δi_a, between points A and B on the curves for V_{a1} and V_{a2} (Fig. 13-11*a*). Divide $\Delta v_a = V_{a1} - V_{a2}$ by Δi_a to obtain r_a corresponding to point Q within the region AB.

When found from anode characteristics (Fig. 13-11*b*), r_a is related to their slope. The greater their slope, the smaller the value of r_a. Thus, the geometric slope of the anode characteristics is connected to $r_a$and not to g_m. The value of r_a is proportional to the cotangent of the angle α that the straight line drawn tangent to the anode characteristic at the specified point Q makes with the anode voltage axis.

It is convenient to find r_a by the two-point method (Fig. 13-11*b*). The value of r_a thus found is the average for the region AB and may be identified with the middle or quiescent point, Q, of the region.

The a.c. anode resistance remains almost constant within the linear portions of the characteristic. Within the tail of the characteristic, r_a increases because of a higher potential barrier and tends to infinity at the cutoff point.

The d.c. resistance R_0 of a vacuum triode is not equal to its a.c. anode resistance, r_a, and can be found in the usual manner from Ohm's law:

$$R_0 = v_a / i_a \qquad (13\text{-}18)$$

The difference between r_a and R_0 is stressed by calling the former the *dynamic anode resistance*, and the latter the *static resistance*. The difference in magnitude between the two may be considerable. R_0 does not remain constant even within the linear regions of the characteristics. It is especially affected by the grid voltage – an increase in the grid voltage brings about a rise in the anode current, so R_0 decreases. The d.c. or static conductance, G_0, of the anode-cathode space depends on the number of electrons residing there. As the grid is made more positive, an ever greater number of electrons fill the space between anode and cathode. Its conductance rises but its resistance falls. An increase in the absolute value of the negative voltage at the grid acts in the reverse direction: the number of electrons in the anode-cathode space decreases while R_0 increases. When the tube is at cutoff, R_0 is equal to infinity.

The *amplification factor* μ and the *penetration factor D* have already been defined, but it is important to discuss them in more detail.

The grid voltage affects the anode current to a greater extent than does the anode voltage. This fact is reflected by the amplification factor μ. The amplification factor tells how many times more effective a change in the grid voltage is than the same change in the anode voltage in controlling the anode current. For example, if the anode voltage needs to be changed by 40 V in order to change the anode current by 1 mA, the equivalent change in the grid voltage will be a mere 2 V. It is clear that the grid is 20 times more effective than the anode, and $\mu = 20$.

Thus, the amplification factor is the ratio of the changes in anode and grid voltages equivalent in terms of their effect on the anode current:

$$\mu = \Delta v_a / \Delta v_g \qquad (13\text{-}19)$$

That is, we take the incremental changes, Δv_a and Δv_g, that produce the same change in anode current, and not just any change.

Let us establish the connection between μ, g_m, and r_a. The transconductance characterizes the effect of grid voltage on anode current, while the quantity characterizing the effect of anode voltage is the a. c. anode conductance $g_a = 1/r_a$. In order to find how many times the grid voltage is more effective than the anode voltage, we must take the ratio of g_m to $1/r_a$; it will yield μ:

$$\mu = g_m/(1/r_a) \quad \text{or} \quad \mu = g_m r_a \qquad (13\text{-}20)$$

This equation, known as the *Barkhausen equation*, is often used in design work. It follows from Eq. (13-20) that if we know any two tube constants or parameters, the third will have the value that will satisfy the above equation. That is, knowing any two tube constants, we can find the third. It should be remembered, however, that if r_a is in ohms, g_m must be expressed in amperes per volt. It is convenient to express r_a in kilohms and g_m in milliamperes per volt. For example, if $g_m = 4$ mA V^{-1} and $r_a = 10$ kΩ, then $\mu = 4 \times 10 = 40$.

Mathematically, the amplification factor of a triode is the absolute value of the ratio of changes in anode and grid voltages that cancel each other, that is, balance out each other's effect on the anode current.

If, for example, an increase of Δv_a in the anode voltage leads to an increase in anode current by Δi_a, the latter can be balanced out by reducing it by the same amount Δi_a. For this purpose, it is necessary to make the grid more negative by Δv_g. Hence, the incremental changes Δv_a and Δv_g that cancel each other must take opposite signs. A negative value of μ has, however, no physical meaning. Therefore, the equation for μ is usually written as

$$\mu = |\Delta v_a / \Delta v_g| \text{ with } i_a \text{ held constant} \qquad (13\text{-}21)$$

or

$$\mu = -\Delta v_a / \Delta v_g \text{ with } i_a \text{ held constant} \qquad (13\text{-}22)$$

The above equations tell us that in order to hold the anode current at a constant value, the anode and grid voltages must be changed in opposite senses and Δv_a should be μ times Δv_g.

The term 'amplification factor' shows how much a tube can boost the applied alternating current. This can best be illustrated by the following example.

Let there be a tube for which $\mu = 10$ and $g_m = 3$ mA V^{-1}. Then feeding an alternating voltage with an amplitude $V_{mg} = 2$ V, will produce an alternating component in the anode current with an amplitude of 6 mA. In other words, a change of 2 V in grid voltage leads to a change of 6 mA in anode current. If we connected in the anode circuit a generator which supplies an alternating emf of 2 V peak value, the change in the anode current would be one-tenth of the above value (because $\mu = 10$), that is, a mere 0.6 mA.

If we wished to obtain in the anode circuit an anode current with an alternating component of 6 mA peak value solely at the expense of an a. c. generator, the latter would have to supply an a. c. emf of 20 V and not 2 V peak value, that is, 10 times greater. Thus, the action of an alternating grid voltage of 2 V peak value is equivalent to connecting in the anode circuit a generator capable of supplying an alternating component of 2 × 10 = 20 V peak value.

Hence, a triode whose grid is fed an alternating voltage V_{mg} may be treated as a generator supplying an alternating emf which is μ times greater, that is, μV_{mg} to the anode circuit. The tube itself, operating as an alternating anode current generator, draws d. c. power from the anode supply source.

Vacuum triodes have an amplification factor μ ranging in value from 3 to 100; most often, it is 10 to 30.

Everything said about μ may be applied to the penetration factor $D = 1/\mu$.

The penetration factor tells how much the action of the anode voltage on the cathode current is attenuated owing to the grid. More accurately, it tells what fraction of the grid's action on the cathode current the anode's action is. Hence, D is defined as

$$D = |\Delta v_g/\Delta v_a| \text{ with } i_k \text{ held constant} \quad (13\text{-}23)$$

or

$$D = -\Delta v_g/\Delta v_a \text{ with } i_k \text{ held constant} \quad (13\text{-}24)$$

As is seen, μ and D are defined subject to different conditions. Therefore, when $i_k > i_a$, the equality $D \approx 1/\mu$ is a very approximate one. It is only when $i_g = 0$ that i_k is precisely equal to i_a.

If we replace μ with D in Eq. (13-20) connecting the three tube parameters, we will have

$$Dr_a g_m = 1 \quad (13\text{-}25)$$

The value of μ (or D) can be found from the tube's characteristics by the two-point method (Fig. 13-12). The procedure is as follows. Referring to the anode-grid curves for V_{a1} and V_{a2} (Fig. 13-12*a*), locate points A and B for the same anode current. The segment AB represents the corresponding change in grid voltage, Δv_g, and the respective change in anode voltage will be $\Delta v_a = V_{a1} - V_{a2}$. Divide Δv_a by Δv_g to obtain μ. The value of μ thus found is the average for the segment AB and is the approximate value of μ at point Q lying midway on the curve (shown dashed) for an anode voltage of $(V_{a1} + V_{a2})/2$.

The amplification factor changes little from one part of a characteristic to another because the horizontal spacing between the curves (segment AB) is nearly constant in magnitude. The fact that changes in μ are small is corroborated by the equation

$$\mu = g_m r_a$$

For example, within the tail of the curve the transconductance g_m decreases while r_a increases in about the same proportion so that their product changes insignificantly. Thus, the amplification factor (or the penetration factor) fits the definition of a tube constant best of all.

When finding the value of μ from the anode characteristics of a tube, locate points A and B for the same value of current on two curves, one for V_{g1} and the other for V_{g2} (Fig. 13-12*b*). The segment AB represents the change in anode voltage, Δv_a. Dividing Δv_a by $\Delta v_g = V_{g1} - V_{g2}$ will give the value of μ which is the average over the segment AB and is close in value to μ at point Q lying midway on the curve for a voltage which is the average of V_{g1} and V_{g2} (the dashed curve).

Everything said about finding μ from the characteristics of a tube fully applies to D.

Figure 13-13 illustrates how one should proceed when finding all the tube parameters for a specified point from the anode characteristics. To begin with, draw a vertical line and a horizontal line through Q-point. The intersections of the two lines with the characteristics will then define g_m (points A and B) and μ (points E and F). The a.c. anode resistance is found from points C and D. A similar procedure can be used to find the tube parameters from a family of anode-grid transfer characteristics.

The manner in which the tube parameters change within the various regions of the characteristics is shown in the plots of g_m, r_a, and μ as functions of grid and anode voltages (Fig. 13-14).

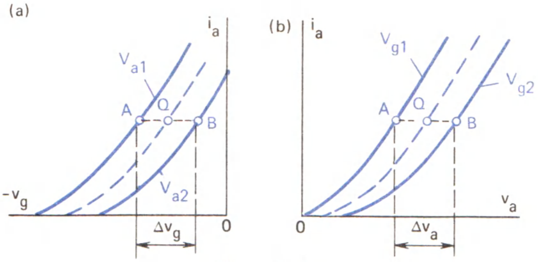


Fig. 13-12

Finding the amplification factor from characteristics

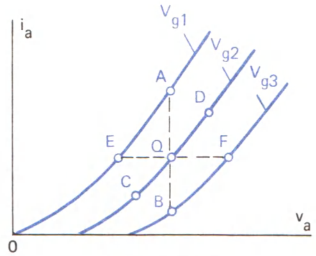


Fig. 13-13

Finding all tube parameters at a specified operating point

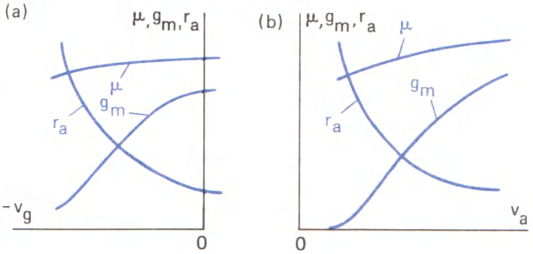


Fig. 13-14

Triode parameters as functions of (*a*) grid voltage and (*b*) anode voltage

Reference sources (data sheets and handbooks) quote the tube parameters for the electrode voltages stated. If a tube is actually operated at some other electrode voltages, the tube constants will be different (especially g_m and r_a). Therefore, when the actual d. c. supply voltages and other operating conditions are different from those stated in a data sheet, the tube parameters will often have to be defined from the tube's characteristics. Inaccuracies in tube manufacture are responsible for a marked spread in parameters among tubes of the same type.

Since the grid-cathode space is equivalent to a diode, investigators are often concerned with the parameters of the diode part of triodes. They are analogous to those of actual vacuum diodes.

13-6 Grid Current

The grid characteristic of a triode may often start at a point other than $v_g = 0$ (curve *2* in Fig. 13-15) because the electrons escaping from the cathode start out on their travel towards the anode at a velocity other than zero, owing to the contact potential difference, and also due to the thermo-emf inevitably generated at a contact between two dissimilar metals and existing in the grid circuit. As often as not, the grid characteristic of a triode starts within the region of small negative grid voltages, or the *tail of the grid characteristics* (curve *1*). The grid current within the tail is very small, being a small fraction of a milliampere for receiving-amplifying tubes. Grid characteristics starting in the region of positive grid voltages (curve *3*) are a rarer occurrence. They result if the contact potential difference turns the grid negative and its action is stronger than that of the initial velocity of the escaping electrons.

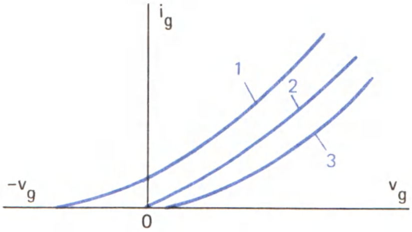


Fig. 13-15

Position of the grid characteristic at several values of grid voltage

Still there is a small grid current even when the grid is negative. It is referred to as the *reverse grid current* (the electrons making up this current move towards the grid in the external leads of the grid circuit). The reverse grid current has three components: an ion current, a thermocurrent, and a leakage current.

The *ion current* occurs in tubes with a poor vacuum (gassy or soft tubes). The grid characteristic of such a tube has the shape shown in Fig. 13-16*a*. When the grid voltage is raised above its cutoff value in the positive direction and there appears an anode current, the electrons travelling towards the anode collide with gas atoms and ionize them. The positive ions

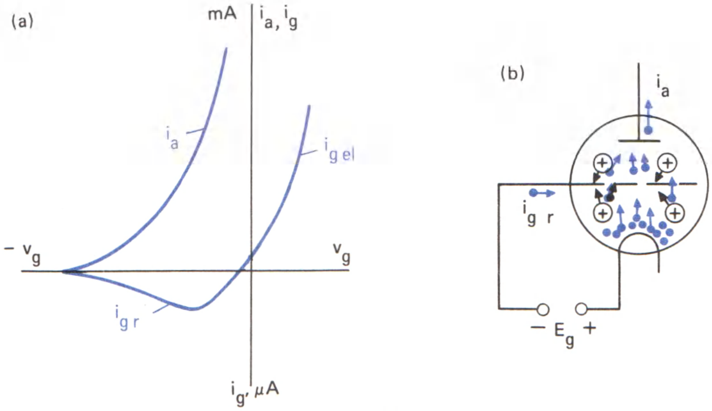


Fig. 13-16

Triode: (*a*) characteristics and (*b*) motion of electrons and ions in a poor vacuum (in a 'gassy' or 'soft' tube) move towards the negative grid and remove electrons from it, thus turning into neutral atoms. The grid expends electrons, but the expenditure is replenished by the grid supply source so that a negative potential is maintained at the grid. The grid circuit carries a current constituted by electrons moving from the "−" terminal of the E_g supply source to the grid (Fig. 13-16*b*).

As the grid is made less negative, the anode current builds up, and ionization is increased. More ions reach the grid, and the ion current increases. As the grid voltage approaches zero, electrons begin to arrive at the grid, giving rise to the electron current of the grid. It rises at a higher rate than the ion current because the number of electrons reaching the grid exceeds that of ions. The reverse grid current decreases until, at some negative voltage, it reverses its direction (Fig. 13-16*a*). From that instant on and with the grid held positive, the electron current prevails over the ion current to the extent that the latter plays no role. As is seen, the reverse grid current reaches its peak value at some negative grid voltage. When the vacuum maintained in the tube envelope is changed, the number of ions is also changed. In consequence, there is a change in the ion current, and the grid characteristic of the tube is shifted. In this way, variation in the vacuum leads to an instability in the characteristics of the tube.

If the grid is raised to an elevated temperature, it may become an emitter of electrons, thus giving rise to a *thermionic emission current* (or *thermocurrent*). As a way of reducing this current, it is practised to make the grid wires in high-power tubes from a metal having a high electronic work function.

The *leakage current* existing in the grid circuit is due to the imperfect insulation between the grid and the other electrodes.

The reverse grid current in high-power tubes does not exceed several microamperes. In low-power tubes, it is a few tenths of a microampere.

Chapter Fourteen
The Vacuum Triode in Operation at Load

14-1 General

A vacuum triode is said to be operating at load, when a load resistance is placed in its anode circuit (Fig. 14-1).

At no-load, the anode voltage of a tube is equal to its anode supply voltage, E_a. If we change grid voltage in this condition, the anode current will change too, but the anode voltage will remain unchanged and equal to E_a. Hence, we may conclude that the anode current is a function of only the grid voltage. In the light of this finding, any calculations related to this condition may be carried out by reference to the usual static characteristics and parameters of the tube. Sometimes, the load is an instrument which has a very low resistance in comparison with the a. c. anode resistance of the tube. Then, it may be deemed that the tube is operating at no-load.

In most cases, however, a tube is operating at load, that is, when the load resistance is comparable with the a. c. anode resistance of the tube. In operation at load, a voltage drop is produced across the load resistance, given by

$$v_R = i_a R_L$$

where R_L is the load resistance. This voltage drop accounts for a noticeable part of E_a. Therefore, we may write

$$v_a = E_a - v_R \text{ or } v_a = E_a - i_a R_L \qquad (14\text{-}1)$$

Equation (14-1) is the basic relation for a tube operating at load.

For simplicity, let us assume that the anode supply source has an internal resistance of zero. Then its voltage is equal to its emf and remains unchanged when the current changes.

Thus, for a tube operating at load we have two conditions: R_L = const and E_a = const. In contrast to operation at no-load, the anode voltage at load does not remain constant. Let, for example, the grid voltage rise so that the anode current increases. Then a greater voltage drop v_R will be produced across the load, while the anode voltage v_a will fall by the same amount because the sum of the two voltages cannot exceed E_a. When the grid voltage decreases the anode voltage must increase.

Thus, in operation at load the anode voltage varies in anti-phase with respect to the grid voltage. Of course, a phase shift of 180° between the anode and grid voltages can result only when the load is purely resistive. If the load includes a reactive component, the phase difference between v_a and v_g will be other than 180°.

Phase reversal of anode voltage relative to grid voltage is responsible for the fact that in operation at load the anode current varies to a lesser degree than it does in operation at no-load. This happens because at no-load the anode current changes solely under the influence of grid voltage; in operation at load a change in anode voltage is in opposition to a change in grid voltage so that the latter change is in part made up for by the opposing change in anode voltage. This is sometimes referred to as *anode reaction.* Of course, the effect of grid voltage is never balanced out completely because the grid produces a stronger action than the anode. Therefore, changes in anode current follow those in grid voltage, that is, are in phase with the latter.

A distinction of a tube operation at load is that its anode current changes in response to the simultaneous and opposing changes in its grid and anode voltages. The anode current is said to be a function of these two voltages:

$$i_a = f(v_g,\ v_a)$$

with the anode voltage itself being a function of the grid voltage in operation at load.

14-2 An Amplifier Stage Using a Vacuum Triode

The most commonly used form of an amplifier stage built around a vacuum triode is the *common-cathode circuit* (Fig. 14-2), similar to the CE transistor stage or the CS FET stage.

The a.c. voltage supplied by a signal source, *SS*, is applied to the grid. The grid-circuit terminals to which the signal voltage is applied are called the *stage input*. Similarly to a transistor amplifier stage, the tube amplifier stage boosts the signal in power. This entails amplification in both voltage and current.

Consider the amplification of a sinewave signal at a relatively low frequency when we may neglect the effect of the interelectrode capacitances.

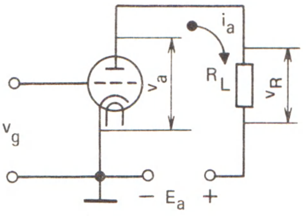


Fig. 14-1

Operation of a triode under dynamic (at load) conditions

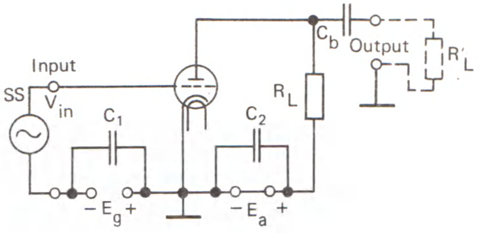


Fig. 14-2

Circuit of an amplifier stage

The source voltage (Fig. 14-3*a*) is given by

$$v_{in} = V_{max} \sin \omega t \tag{14-2}$$

The grid also accepts a negative direct voltage, E_g, called the *grid bias voltage.* It shifts the operation of the tube into the region of negative grid voltages and has as its primary objective to prevent the grid from drawing any current. Otherwise this would cause distortion in the amplified signal and put an additional load on the signal source with the result that the alternating voltage applied to the grid would be reduced. So long as the bias voltage E_g is not smaller in absolute value than the amplitude of the signal being amplified, V_{mg}, the grid remains always negative and draws no current. Thus, if we wish to avoid a grid current, we should meet the following condition:

$$|E_g| \geqslant V_{mg} \tag{14-3}$$

The resultant grid voltage is a pulsating one (Fig. 14-3*b*) and is given by

$$v_g = E_g + V_{mg} \sin \omega t \tag{14-4}$$

where $V_{mg} = V_{m,\text{in}}$.

This voltage causes the anode current to pulsate as well. When there is no alternating voltage applied to the grid (the *no-signal* or *quiescent state*), the stage is quiescent, and the anode current has a constant value, I_{a0}, called the *quiescent anode current*. When an a.c. voltage is fed to the grid (the *with-signal state*), it causes the current to change periodically. If the tube is operating within the linear region of its anode-grid characteristic, the anode current varies in the same manner as the grid voltage. As a result, the anode current acquires a sinewave component of amplitude I_{ma} (Fig. 14-3*c*). The anode current is then given by an equation of the form

$$i_a = I_{a0} + I_{ma} \sin \omega t \qquad (14\text{-}5)$$

This current produces across the load resistance R_L a voltage drop, $v_R = i_a R_L$, which changes in the same manner as the anode current. Therefore, the plot of the anode current waveform, drawn to a different scale, may represent the waveform of v_R:

$$v_R = V_{R0} + V_{mR} \sin \omega t \qquad (14\text{-}6)$$

where

$$V_{R0} = I_{a0} R_L$$

and

$$V_{mR} = V_{m\,out} = I_{ma} R_L \qquad (14\text{-}7)$$

The anode voltage varies in anti-phase with respect to v_g and i_a (Fig. 14-3*d*). In the quiescent state,

$$V_{a0} = E_a - V_{R0}$$

In the with-signal state, the anode voltage is given by

$$v_a = V_{a0} - V_{ma} \sin \omega t \qquad (14\text{-}8)$$

If we move around the anode circuit consecutively, we will see that v_a and v_R change in opposite senses in accord with Eq. (14-1). The alternating voltages at the anode and across the load, as found with respect to the cathode, will be equal, that is,

$$V_{ma} = V_{mR}$$

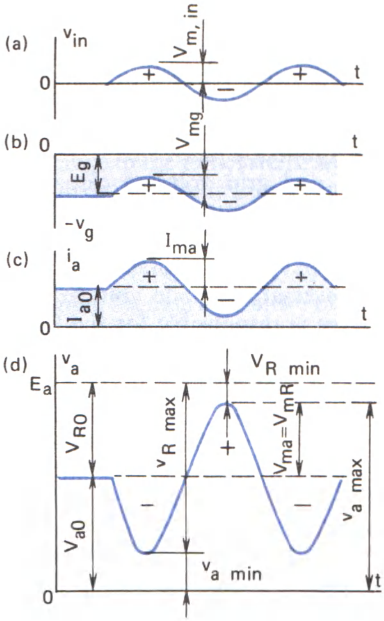


Fig. 14-3

Operation of an amplifier stage built around a triode

Thus, the output voltage is the alternating anode voltage. Therefore, the anode and cathode of the tube are its output terminals. When it is required that no d.c. voltage should be present at the output, a d.c. blocking capacitor, C_b, is placed between the anode and the output terminal (see Fig. 14-2). It passes only the amplified alternating voltage, but acts as an effective open-circuit for the d.c. component. The capacitance of the d.c. blocking capacitor is chosen such that its reactance at the lowest operating frequency is a small fraction of the load resistance, R_L', connected to the output terminals. Then the loss of alternating voltage across C_b will be negligible. If the reactance of the blocking capacitor, $1/\omega C_b$, is less than $0.1\ R_L'$, the reduction in the output voltage due to C_b will be less than 0.5%.

The E_a supply source is shunted by a capacitor, C_2, whose reactance at the lowest operating frequency is a small fraction of R_L. Its purpose is to eliminate the effect of the internal resistance of the anode supply source because the low reactance of a capacitor is equivalent to a short-circuit for an alternating current, and no a.c. voltage drop occurs across the capacitor. As often as not, C_2 is not shown in the circuit diagram, implying that it is lumped with the E_a supply source. For example, a rectifier always includes a high-value capacitor to smoothen the ripple. The grid bias voltage source is likewise shunted by a capacitor, C_1.

Amplifier stages often use *automatic bias* (also called self-bias). The required magnitude of bias voltage is developed from the E_a supply source

(Fig. 14-4) by means of R_k, called the *cathode resistor* or the *self-bias resistor*, placed in the cathode lead and bypassed by a capacitor, C_k. The direct component of cathode current produces across R_k a voltage drop which is utilized as the grid bias voltage

$$E_g = I_{k0} R_k \tag{14-9}$$

The positive side of this voltage is applied to the cathode and its negative side (via the signal source *SS* or a grid resistor R_g) to the grid. Using Eq. (14-9), we can find the value of R_k required to develop the necessary bias voltage. For example, if we wish to obtain $E_g = -4$ V at $I_{k0} = 5$ mA, then

$$R_k = E_g/I_{k0} = 4 \div 5 = 0.8 \text{ k}\Omega = 800\ \Omega$$

The cathode capacitor C_k has a sufficiently high capacitance and removes the ripple from the voltage across R_k due to the alternating component of cathode current. This capacitor acts in a manner similar to that of the capacitor intended to smoothen the ripple in a rectifier (see Chap. 3). For the ripple in the bias voltage to be kept to a minimum the reactance of C_k at the lowest operating frequency ω_l must be a small fraction of R_k:

$$1/\omega_l C_k \ll R_k \tag{14-10}$$

Then it is legitimate to deem that the alternating cathode current is diverted to pass through C_k, and only the direct component of cathode current is able to flow through R_k.

The circuit shown in Fig. 14-4*a* is used when the signal source *SS* does not present an open-circuit to direct current and when the signal source itself does not generate a direct voltage. If, on the other hand, the signal source conducts no direct current or generates a direct voltage, resort is made to the circuit shown in Fig. 14-4*b*. Here, the signal to be amplified is fed to the grid via a d.c. blocking capacitor C_b, and the bias voltage is fed via R_g, a grid resistor of a very high value (usually hundreds of kilohms or even greater) so as to make the input resistance of the stage very high. To avoid the loss of alternating current, the capacitive reactance of C_b at the lowest operating frequency ω_l must be a small fraction of R_g:

$$1/\omega_l C_b \ll R_g \tag{14-11}$$

This resistor also serves to prevent an excessive accumulation of electrons at the grid. If there were no R_g, the grid circuit would be open for direct current (the grid would be isolated), and the electrons reaching the grid would charge it to a negative potential such that the tube would be driven to cutoff. With R_g included in the grid circuit, the charge accumulating on the grid is free to leak to ground. This is the reason why R_g is often called the *grid leak*.

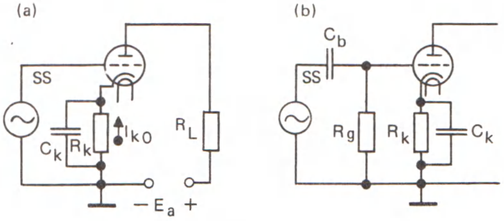


Fig. 14-4
Automatic (self-) bias circuits

If the grid leak is not to overload the signal source, it must have a sufficiently high value, usually such that $R_g \gg R_{ss}$. Because the input alternating voltage is divided between C_b and R_g, it is essential to satisfy the condition defined in Eq. (14-11). Then the loss of voltage across C_b will be insignificant, but R_g ought not to be chosen too high in value. When a large pulse of positive voltage (say, due to noise or interference) happens to reach the grid, the grid attracts a great number of electrons and accumulates a large negative charge which tends to drive the tube to cutoff. At a very high value of R_g, this charge will leak off very slowly, and the tube will remain in the OFF (nonconducting) state for some time.

Now let us see why the grid current is a detrimental factor. Suppose that an amplifying stage is operating with no negative grid bias applied. Then, during the negative half-cycles of the alternating grid voltage, no current will be flowing in the grid circuit, the signal source will be idling, and the grid voltage will be equal to the signal source emf. During the positive half-cycles, however, a grid current will be flowing, and it will produce a voltage drop across the internal resistance of the signal source, R_{ss}. Now the signal source will be operating at load, and the grid voltage will be smaller than its emf by the fraction of voltage lost inside the signal source. During the positive half-cycles, the amplitude of grid voltage will be

$$V'_{mg} = E_{mg} - I_{g\,\max} R_{ss} \tag{14-12}$$

where E_{mg} = amplitude of the signal-source emf

$I_{g\,max}$ = maximum value of grid current

During the negative half-cycles, $V''_{mg} = E_{mg}$.

The internal resistance of the signal source, R_{ss}, is often very large. The voltage drop across the signal source is then appreciable, too. As a result, the alternating grid voltage ceases to be sinewave, that is, nonlinear distortion occurs. The peak value of grid voltage during the positive half-cycles is smaller than it is during the negative half-cycles (Fig. 14-5*a*). Also, an increase in the amplitude of alternating grid voltage leads to a greater grid current and a heavier distortion. It is caused by the nonlinearity of R_{gk}, the resistance of the grid-cathode space which acts similarly to a vacuum diode. When the grid is positive, this resistance does not exceed 1 kΩ. When the grid is negative, it tends to infinity. When the signal source is loaded into such a nonlinear resistance, its voltage waveform is distorted. Since the grid voltage waveform is distorted, the amplified voltage appearing at the output of the amplifying stage likewise has a distorted waveform.

Nonlinear distortion is at its heaviest when R_{ss} is many times R_{gk}. Then, during the positive half-cycles, the signal source is short-circuited, and the grid voltage is zero very nearly. For example, if R_{ss} = 100 kΩ and R_{gk} = 1 kΩ, then the grid voltage during the negative half-cycles is equal to the emf of the signal source, while during the positive half-cycles it is about 1% of the signal-source emf. That is, nearly all of the signal-source emf is dropped across R_{ss} (Fig. 14-5*b*). Practically, the voltage fully consists of only negative half-cycles (the positive half-cycles are clipped off).

This condition is utilized in limiters, but it cannot be used for distortion-free amplification.

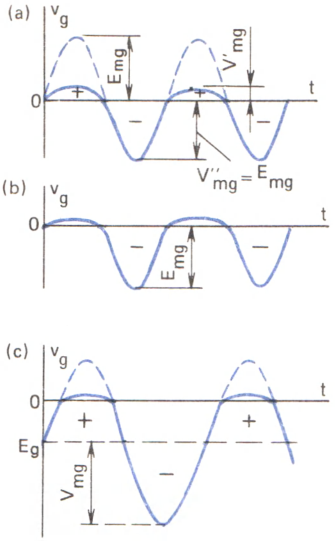


Fig. 14-5

Grid voltage of an amplifier stage at various bias values

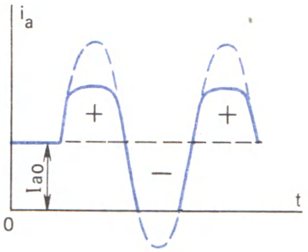


Fig. 14-6

Anode current distortion in the case of positive- and negative-peak cut-off

When the grid is prevented from drawing any current in operation by applying a negative grid bias, the resistance of the grid-cathode space is very high, and the signal source is operating at no-load for the duration of each cycle. The grid voltage is at all times equal to the emf of the signal source, and its peak value is the same during both the positive and negative half-cycles and has the maximum attainable magnitude. In the circumstances, the grid-cathode space does not load the signal source, that is, it draws no power from it. In consequence, we may use a signal source of an arbitrary low power rating. There is some load on the source due to the input capacitance of the tube, but it is negligible and has to be taken into account only at very high frequencies (see Sec. 14-7).

Thus, when an amplifier stage is operating with its grid held negative so that it draws no current, no distortion can occur that would otherwise be caused by this current. If, however, the alternating grid voltage exceeds in amplitude the grid bias voltage, $V_{mg} > |E_g|$, there will be a grid current flowing during some part of each cycle, and it will give rise to distortion. The part of the positive half-cycle of alternating grid voltage that extends into the positive region will be clipped off (Fig. 14-5*c*).

The waveform of plate current will be distorted as well. As is shown in Fig. 14-6, the upper part of the positive cycle of anode current is clipped off or flattened (*positive-peak limiting*). If a part of the negative half-cycle of grid voltage happens to drive the tube to cutoff, it is clipped off too (*negative-peak limiting*). Instead of a sinewave, the anode current waveform may become trapezoidal.

When a d. c. voltage exists in the signal source, it ought not to be allowed to reach the grid. To this end, the signal to be amplified is fed to the tube via a transformer (Fig. 14-7) or a d.c. blocking capacitor (Fig. 14-4*b*). The bias voltage is applied to the grid via the transformer secondary or a grid resistor, R_g, with a resistance of several hundred kilohms to several megohms.

Tube-based amplifier stages may be used at any frequencies. At audio frequencies, it is customary to use *resistance-* (or *resistive-*) *coupled stages* (see Fig. 14-4*a*), *transformer-coupled stages* (Fig. 14-8*a*), and, although more seldom, *choke-coupled stages* (Fig. 14-8*b*).

Radio-frequency (r. f.) amplifiers include, as a rule, a resonant circuit. The input section of such stages may have any of the configurations we have examined: the signal source may be directly coupled to the tube (Fig. 14-9*a*), transformer-coupled (see Fig. 14-7), or capacitance-coupled (see Fig. 14-4*b*). As often as not, an r. f. amplifier may have a second resonant circuit in its grid lead (Fig. 14-9*b*).

14-3 Amplifier Stage Parameters

The key parameter is the *stage gain* k_V defined as the ratio of output to input voltage:

$$k_V = V_{m\,\mathrm{out}}/V_{m,\mathrm{in}} = V_{mR}/V_{mg} \qquad (14\text{-}13)$$

Although k_V is the *stage voltage gain*, the word 'voltage' is often omitted.

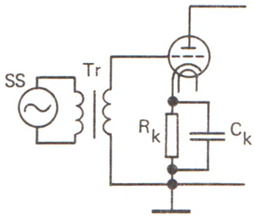


Fig. 14-7
Input coupling transformer

The amplification of alternating current is expressed in terms of the *stage current gain* k_I defined as the ratio between the peak values of output and input currents:

$$k_I = I_{m\,\mathrm{out}}/I_{m,\mathrm{in}} = I_{ma}/I_{mg} \qquad (14\text{-}14)$$

If a stage is operating at low frequencies and its grid is drawing no current, the input current is negligible, and k_I is very high. It may be several thousands or even millions, that is, many times the figure for stages using bipolar transistors.

The amplified voltage appearing at the stage output is given by

$$V_{m\,\mathrm{out}} = V_{mR} = V_{ma} = I_{ma}R_L$$

or $\qquad (14\text{-}15)$

$$V_{mR} = KV_{mg}$$

The performance of an amplifier stage can also be stated in terms of its *output power*, P_{out}, which is the power due to the alternating component of anode current in the load:

$$P_{\mathrm{out}} = 0.5I_{ma}V_{mR} = 0.5I_{ma}^2R_L = 0.5V_{mR}^2/R_L(14\text{-}16)$$

The efficiency of an amplifier stage may be expressed in any one of several ways. One is the *anode circuit efficiency* defined as the ratio between its output power P_{out} and the d.c. power P_0 supplied by the E_a source:

$$\eta = P_{\mathrm{out}}/P_0 \qquad (14\text{-}17)$$

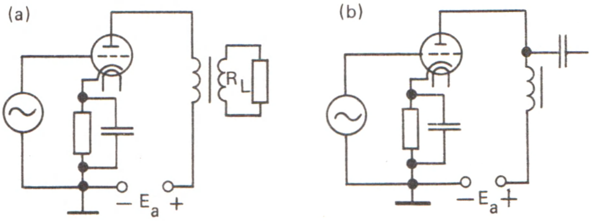


Fig. 14-8
Amplifier stage: (*a*) transformer-coupled and (*b*) choke-coupled

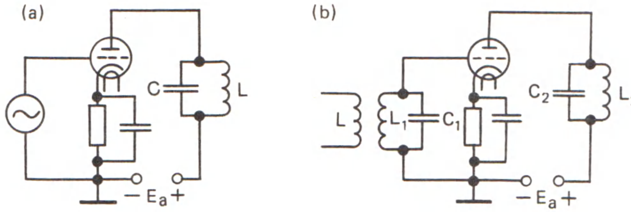


Fig. 14-9
Tank-circuit-coupled amplifier stages

where

$$P_0 = I_{\mathrm{a\,av}} E_{\mathrm{a}} \qquad (14\text{-}18)$$

or the product of the anode source emf E_{a} by the d. c. component of anode current, $I_{\mathrm{a\,av}}$.

The efficiency tells what fraction of the power expended by the anode supply source is converted to the useful power of the amplified signal.

The difference between P_0 and P_{out} is the *power loss*:

$$P_{\mathrm{loss}} = P_0 - P_{\mathrm{out}} \qquad (14\text{-}19)$$

In a resistive-coupled stage, the power loss is the sum of the anode dissipation P_{a} and the d. c. power lost in the load resistor, P_{R0}. The efficiency of resistive-coupled stages is always low, but this factor is of minor importance because such stages are used as low-power amplifiers. High efficiency becomes an important factor where high power is involved. The efficiency of high-power transformer-coupled a. f. amplifiers or tuned-circuit r. f. amplifiers operating at low distortion levels is 40-45%. These circuits have a higher efficiency because, among other things, the d. c. resistance of the transformer primary or the tuned-circuit inductor is small and they dissipate very little power. For these stages, the power loss is approximately equal to the anode dissipation alone:

$$P_{\mathrm{loss}} \approx P_{\mathrm{a}} = P_0 - P_{\mathrm{out}} \qquad (14\text{-}20)$$

When no a. c. voltage is fed to the grid and $P_{\mathrm{out}} = 0$, all of P_0 is equal to P_{a}, that is, it is dissipated at the anode. This might cause overheating, and the tube might fail.

In high-power stages operating in modes that permit substantial distortion but calling for the largest possible power output and efficiency, the latter may be as high as 70-80%.

Efficiency can be improved by the use of negative grid bias. It reduces the direct component of anode current and, in consequence, P_0. An excessive negative bias, however, reduces the useful power output because the tube transconductance falls and appreciable nonlinear distortion arises.

Stage efficiency ought not to be confused with the *stage power gain*, k_P, defined as the ratio of the output power, P_{out}, to the signal-source power, P_{in}:

$$k_P = P_{\mathrm{out}}/P_{\mathrm{in}} \qquad (14\text{-}21)$$

The signal-source power can be found by the equation

$$P_{\mathrm{in}} = 0.5 I_{m,\mathrm{in}}/V_{m,\mathrm{in}} \qquad (14\text{-}22)$$

Therefore,

$$k_P = 0.5 I_{m\,\mathrm{out}} V_{m\,\mathrm{out}}/0.5 I_{m,\mathrm{in}} V_{m,\mathrm{in}} = k_I k_V \qquad (14\text{-}23)$$

A major distinction of an amplifier stage is that it boosts the input signal in power. How well it does this job is indicated by its k_P. There is some difficulty, however, in finding P_{in}. Therefore, one ordinarily uses only the stage voltage gain k_V. For an a. f. amplifier stage operating at negative grid bias, P_{in} is negligible because the grid current is very small. If the stage includes, in addition, a grid leak R_{g} (see Fig. 14-4*b*), P_{in} will be determined by the power lost in the grid resistor:

$$P_{\mathrm{in}} = V_{mg}^2/2R_{\mathrm{g}} \qquad (14\text{-}24)$$

Since R_{g} is usually very high, the input power is negligible. For example, when $V_{mg} = 2$ V and $R_{\mathrm{g}} = 1$ mΩ, we have

$$P_{\mathrm{in}} = 2^2/(2 \times 10^6) = 2 \times 10^{-6}\ \mathrm{W} = 2\ \mu\mathrm{W} \qquad (14\text{-}25)$$

In amplifiers operating with no grid current, k_P may be as high as several hundred thousands or even greater. It is always much lower for stages using bipolar transistors owing to appreciable input currents. When an amplifier is operating with its grid drawing some current, there is an abrupt fall in its k_P because of a substantial increase in P_{in}. There is a reduction in k_p also at microwave frequencies (see Chap. 19). In the UHF and SHF bands it may even fall to less than unity.

Another important quantity of an amplifier stage is its *input impedance*, that is, the opposition that the stage presents to the signal source, R_{in}. In fact, we should have given it the symbol Z_{in} because it has both a resistive and a reactive component. The reactive component is contributed by the input capacitance of the tube (see Sec. 14-8). At audio frequencies, however, this reactive component is very high, so it is legitimate to neglect its shunting effect and to regard the input impedance as purely resistive; hence its designation as R_{in}. In r. f. amplifiers, the grid circuit is usually extended to include a resonant circuit, so the input capacitance is lumped with the tuned-circuit capacitance, and again we need only to consider the resistive component of the input impedance, that is, R_{in}.

With no grid current flowing and at low frequencies, R_{in} tends to infinity. Practically, this resistance may be very high (many megohms), which is an advantage because the signal source is then operating in conditions close to an open circuit (at no load), and it delivers a maximum voltage nearly equal to its emf. If, in addition, there is a grid leak, R_g, as shown in Fig. 14-4*b*, the input resistance will solely be determined by its value. The grid current drastically reduces the input resistance to a few kilohms or several hundred ohms. If the signal source generates a sinewave emf, the input resistance is that of the grid-cathode space at the fundamental of the pulsating grid current:

$$R_{\text{in}} = V_{mg1}/I_{mg1} \tag{14-26}$$

where V_{mg1} = fundamental amplitude of grid voltage
I_{mg1} = fundamental amplitude of grid current

The performance of the tube used in an amplifier stage can be calculated using an analytical method or a grapho-analytical method.

With the *analytical method*, calculations involve the use of simple equations written in terms of the tube parameters usually found from the characteristics applicable to a given set of operating conditions. The ratings quoted in data sheets and other reference sources may only be used when the actual operating conditions are close to the standard ones. Unfortunately, this method is not accurate enough in cases where the signal to be amplified has large peak values. The point is that it does not take into account the tube nonlinearity. Nor may the analytical equations be used to calculate the direct components of tube currents and voltages.

The *grapho-analytical method* is based on the use of the load line and dynamic transfer characteristics which are derived from the static characteristics and take into account the tube nonlinearity. Therefore, this method is most accurate. Also, it permits the calculation of both alternating and direct components of currents and voltages.

Unfortunately, the graphical constructions involved are very time-consuming. Also, the method is not applicable to small-signal cases when the necessary graphical construction cannot simply be done, and calculations have to be carried out using equations written in terms of the tube parameters.

14-4 The Analytical Method and the Equivalent Circuits of an Amplifier Stage

This method uses the tube parameters in order to determine the incremental changes in anode current, Δi_a, in response to changes in v_g and v_a.

Because in operation at load changes occur in both the grid and anode voltages, Δi_a may be interpreted as a sum of two increments: $\Delta i_a'$ due to Δv_g while neglecting the anode reaction, and $\Delta i_a''$ due to changes in anode voltage by Δv_a.

It follows from the equation of transconductance that

$$\Delta i_a' = g_m \Delta v_g \tag{14-27}$$

Recalling the equation for r_a, we may write

$$\Delta i_a'' = \Delta v_a / r_a \tag{14-28}$$

Hence, the total incremental change in anode current is

$$\Delta i_a = \Delta i_a' + \Delta i_a'' \tag{14-29}$$

or

$$\Delta i_a = g_m \Delta v_g + \Delta v_a / r_a \tag{14-30}$$

Equation (14-30) is the *basic tube equation*. Using it, we can, among other things, derive an equation connecting the tube constants. To demonstrate, if $\Delta i_a = 0$ or, which is the same, $i_a = \text{const}$, we get

$$g_m r_a = -\Delta v_a/\Delta v_g = \mu$$

Let us re-write Eq. (14-30) in a more convenient form. A change in anode voltage is always equal in magnitude but opposite in sign to the corresponding change in the voltage across the load resistance R_L:

$$\Delta v_a = -\Delta v_R \qquad (14\text{-}31)$$

By Ohm's law, Δv_R is equal to $R_L \Delta i_a$. Hence,

$$\Delta v_a = -R_L \Delta i_a^* \qquad (14\text{-}32)$$

On substituting Eq. (14-32) into Eq. (14-30), we get

$$\Delta i_a = g_m \Delta v_g - R_L \Delta i_a / r_a \qquad (14\text{-}33)$$

Solving Eq. (14-33) for Δi_a yields

$$\Delta i_a = g_m r_a \Delta v_g/(r_a + R_L) \qquad (14\text{-}34)$$

or (if we recall that $g_m r_a = \mu$)

$$\Delta i_a = \mu \Delta v_g/(r_a + R_L) \qquad (14\text{-}35)$$

Equation (14-35) is the basic one in the analysis and synthesis of tube circuits and expresses *Ohm's law for alternating anode current*. The numerator $\mu\Delta v_g$ is the alternating emf existing in the anode circuit, and the denominator $r_a + R_L$ is the total anode circuit resistance to alternating current. It follows then that in the anode circuit the tube acts as a generator of an alternating emf (terminal voltage) $\mu\Delta v_g$. The same conclusion was drawn in connection with the amplification factor μ. Now we have obtained it mathematically.

Thus, a triode whose grid voltage changes by Δv_g is equivalent to an ideal generator of an alternating emf (terminal voltage) $\mu\Delta v_g$ and with an internal resistance r_a. Of course, a tube can operate as a voltage generator only on the proviso that its anode circuit is energized by a d. c. emf source and an a. c. voltage is fed to its grid.

In terms of alternating current, the anode circuit of a triode may be represented by the equivalent circuit of Fig. 14-10*a*. It does not include the anode supply source because its resistance to the alternating component is deemed equal to zero. Sometimes, the equivalent circuit shows the ideal $\mu\Delta v_g$ generator in series with a resistor, r_a (Fig. 14-10*b*). The $\mu\Delta v_g$ generator is the tube. The anode supply source feeds a d. c. emf (terminal voltage), E_a. Its purpose is to supply the anode circuit with direct voltage. The load R_L dissipates and not generates power. The alternating component of anode current is produced solely inside the tube when a change in its grid voltage, Δv_g, causes its anode current to vary.

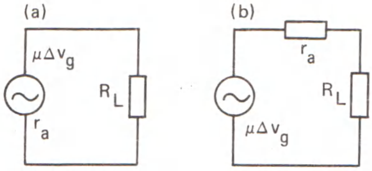


Fig. 14-10

Equivalent anode circuit for the a. c. component of anode current, with the triode replaced by a Thévenin equivalent (a constant-voltage generator)

The treatment of a tube as an alternating voltage (or, rather, emf) generator was proposed independently of each other by Bonch-Bruyevich of Russia and H. Barkhausen of Germany. Equation (14-35) and the associated equivalent circuit have proved very useful in tube circuit analysis and design. A good deal of theory dealing with tube and other electronic circuits has been based on this classical concept. Its opponents argue, however, that a tube cannot be treated as a voltage generator. They neglect the fact that a generator is an energy (or power) converter. It uses one form of energy and generates another form of energy. In our case, we feed d. c. energy from the anode supply source to a tube which converts some of the input into an a. c. energy. Therefore, the treatment of a tube as an a. c. voltage generator has a real physical meaning. It is in the tube that the a. c. emf giving rise to an a. c. anode current is generated.

The opponents of the classical theory treat the tube as a variable resistor for which the equivalent circuit should take the form shown in Fig. 14-11. This equivalent circuit is physically sound, too, and applies to both the alternating and the direct components of anode current. If the grid voltage is constant, the tube presents a certain static (or d. c.) resistance, R_0, and the anode current is

$$I_{a0} = E_a/(R_0 + R_L) \qquad (14\text{-}36)$$

[*] This result can also be obtained on finding Δv_a from the equation: $v_a = E_a - i_a R_L$.

When the grid becomes more negative, R_0 increases. When it becomes more positive, R_0 decreases. If, however, R_0 changes, the anode current must change as well–it acquires an alternating component.

However, it is not convenient to use the equivalent circuit of Fig. 14-11 for practical purposes. The point is that while the anode current varies sinusoidally (such as when the tube is operating at a small sinewave alternating grid voltage), R_0 has to vary in an intricate, nonsinusoidal manner, and this fact entails considerable mathematical handicaps.

The equivalent circuit for an alternating anode current, with the tube replaced by an equivalent voltage generator, is simple and convenient to use. Ohm's law, Eq. (14-35), applicable to this equivalent circuit establishes a linear relationship between changes in anode current and changes in grid voltage, Δv_g. So long as the grid voltage varies sinusoidally, the anode current likewise changes sinusoidally. This is the reason why the equivalent circuit of Fig. 14-10 has won wide popularity despite the fact that it cannot be applied to direct anode current.

Equation (14-35) yields accurate results only when the tube is operated within the linear regions of its characteristics where μ and r_a are constant. Within the nonlinear regions of the characteristics, μ and r_a themselves are functions of grid voltage. If, with the tube operating within the nonlinear regions of its characteristics, we use Eq. (14-35) and substitute in it the values of μ and r_a averaged over those regions, the results will be approximate. The error can be reduced by reducing the change in the grid voltage, Δv_g. Equation (14-35) is a *linear approximation* of the nonlinear relationship between the changes in anode current and grid voltage. It may also be used to find their peak values:

$$I_{ma} = \mu V_{mg}/(r_a + R_L) \qquad (14\text{-}37)$$

The smaller the amplitude, the smaller the error.

After the amplitude of the alternating component of anode current has been found, it is an easy matter to determine the output voltage and the output power of the stage.

Sometimes it is convenient to treat a tube as an equivalent current generator (or constant-current source) with the help of Norton's theorem. By Norton's theorem, any voltage generator whose terminal voltage (or emf) is E may be replaced with an equivalent constant-current generator which produces a short-circuit current equal to the quotient of its terminal voltage by its internal resistance, the latter being connected in parallel with the load resistance. This type of equivalent circuit is shown in Fig. 14-12. Here, the alternating current Δi_a is flowing, as before, through R_L, and the generator current given by $g_m \Delta v_g$ is the short-circuit current, that is, the current at no-load. To demonstrate, it follows from Eq. (14-35) that when $R_L = 0$, the change in current is equal to

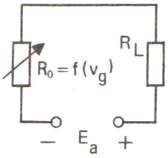


Fig. 14-11

Equivalent anode circuit, with the triode replaced by a variable resistor

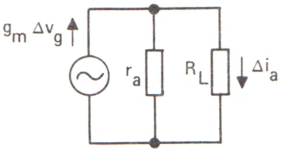


Fig. 14-12

Equivalent anode circuit, with the triode replaced by a Norton equivalent (a constant-current generator)

$$\mu \Delta v_g / r_a = g_m \Delta v_g$$

Let us prove the validity of the Norton equivalent circuit. To this end, we multiply both sides of Eq. (14-34) by R_L:

$$R_L \Delta i_a = g_m \Delta v_g r_a R_L/(r_a + R_L) \qquad (14\text{-}38)$$

The product of R_L by Δi_a is the voltage Δv_R, and the right-hand side of the equation shows that Δv_R can be obtained if we multiply the current $g_m \Delta v_g$ by the total resistance of r_a and R_L connected in parallel. However, the parallel connection of the internal resistance of the generator and the load is a convention which does not always reflect the true situation. The Norton equivalent circuit is more often used in cases where the load is a parallel combination of several arms and also in connection with pentode tubes.

Now we will go through a procedure that yields the *stage voltage gain* which is defined as

$$k_V = \Delta v_R / \Delta v_g \qquad (14\text{-}39)$$

The stage gain is physically different from the amplification factor of a tube. The amplification factor of a tube is defined as the absolute ratio of increments in anode and grid voltages such that the anode current of the tube remains constant (see Sec. 13.5). In Eq. (14-39), the change in load voltage, Δv_R, results from a change in grid voltage, Δv_g. In other words, the stage gain k_V tells how many times the alternating voltage applied to the tube's input is amplified.

Since $\Delta v_R = R_L \Delta i_a$, it follows that

$$k_V = R_L \Delta i_a / \Delta v_g \qquad (14\text{-}40)$$

If we replace Δi_a in Eq. (14-40) with its expression from Eq. (14-35), and cancel out Δv_g, we will derive a very important formula

$$k_V = \mu R_L / (r_a + R_L) \qquad (14\text{-}41)$$

Equation (14-41) is widely used in telecommunications and electronics. If we know the tube constants and the load resistance, we can readily calculate the stage voltage gain by this equation. As often, it may be used to solve an inverse problem, that is, to find R_L at which a given tube would yield the necessary gain. It is seen from Eq. (14-41) that in operation at load k_V is less than μ because it is multiplied on the right-hand side of the equation by a fraction which is smaller than unity. This means that the alternating emf, $\mu \Delta v_g$, generated by the tube cannot be fully utilized. The greater the load resistance in comparison with r_a, the greater the contribution of Δv_R to the alternating emf and the closer the value of k_V to that of μ. The fraction $R_L/(r_a + R_L)$ increases and tends to unity with increasing R_L.

Example. Suppose that we have a tube for which $\mu = 10$ and $r_a = 10$ kΩ, and that the load resistance is $R_L = 40$ kΩ. Then, by Eq. (14-41), we obtain

$$k_V = 10 \times 40/(10 + 40) = 8$$

so k_V is less than μ. If the alternating voltage fed to the grid is $\Delta v_g = 2$ V, the alternating emf acting in the anode circuit will be $\mu \Delta v_g = 10 \times \times 2 = 20$ V. It will be divided between R_L and r_a. The share of R_L will be 16 V, or 80% of the emf, because 40 kΩ is 80% of the total resistance of the anode circuit, equal to 50 kΩ. Thus,

$$k_V = 16/2 = 8$$

Suppose now that R_L tends to infinity. Then, by Eq. (14-41), we find that k_V seems to tend to μ.* This is not feasible, however, because at $R_L = \infty$, the anode circuit is open, and the tube will not operate.

As R_L is increased, the stage gain k_V first rises at a high rate, then its increase slows down on approaching μ. When a triode is used in a voltage amplifier, R_L is chosen to be not more than 3 or 4 times r_a because any further increase in R_L would not lead to a marked increase in amplification. It must also be remembered that some of the direct voltage supplied by the anode source is lost across R_L. When R_L is too high, the anode voltage falls too much, and the tube is forced to operate within the tail of its characteristics where μ is low but r_a is high (see Fig. 13-14*b*). This leads to a reduced value of k_V. In practice, it is usual to choose for triodes

$$R_L = (1 \text{ to } 4)\, r_a \qquad (14\text{-}42)$$

in which case k_V ranges from 0.5 to 0.8 of μ.

14-5 Graphical Analysis of Triode Performance at Load

The graphical analysis of tube performance at load (that is, when a tube is operating as a voltage amplifier) is based on the *load* or *dynamic characteristics* that can be constructed on a family of static characteristics if we know the anode supply voltage E_a and the load resistance R_L. The simplest and most accurate way to do this is with the aid of what is called the *load line*. In order to construct it, one needs a family of anode characteristics (Fig. 14-13). The load line is described by the equation

$$v_a = E_a - i_a R_L$$

When v_a is plotted as abscissa and i_a as ordinate, this first-degree equation yields a straight line. It can conveniently be constructed by two points. Letting $i_a = 0$ gives $v_a = E_a$, and this locates point M on the axis of abscissas. At this point, the tube is driven to cutoff by a negative grid voltage. If the tube is at cutoff, and there is no anode current flowing, no voltage drop can be produced across R_L, and all of the E_a supply voltage is applied to the tube.

To locate the second point, we set $v_a = 0$. This

* Since the substitution in Eq. (14-41) results in an indeterminate form, ∞/∞, the numerator and denominator must first be divided by R_L.

yields $i_a = E_a/R_L$. On the plot it appears as point N. Now we draw a straight line through points M and N; this is the load line. It is to be noted that point N does not represent any real operating conditions of the tube. At $v_a = 0$, the anode current cannot be a maximum.

Using the load line, we can find the anode current and the anode voltage at any grid voltage. As an example, Fig. 14-13 shows that at V_{g3} the values of i_a and v_a are represented by point B. The intercept completing v_a to E_a represents the voltage drop across the load, v_R. We can also find the grid voltage that corresponds to the desired anode current.

The greater the value of R_L, the smaller the slope of the load line. When $R_L = 0$, the load line runs vertically (line MB' in the drawing). This corresponds to operation at no-load when $v_a = E_a = \text{const}$. It is seen from Fig. 14-13 that in operation at no-load and at a grid voltage V_{g3}, the anode current is represented by point B', while in operation at load it is smaller (point B) because the anode voltage is brought down by the voltage drop across the load, v_R. At $R_L = \infty$, the load line merges with the axis of abscissas, and the anode current is zero at any value of voltage.

In addition to the values of E_a and R_L, which are required in order to construct the load line, we need to know the grid bias E_g and the peak value of the alternating component of grid voltage, V_{mg}. They may be specified or chosen from some considerations. For example, if the objective is to minimize signal distortion, then E_g and V_{mg} must be such that the tube is operating within the linear regions of the characteristics and with its grid drawing no current. Figure 14-14 shows the construction for a more general case of amplification with some distortion due to the nonlinear region of the characteristics. The bias voltage E_g determines the position of the *quiescent* (Q) *point* on the characteristic, the quiescent plate voltage V_{a0}, the quiescent direct anode current I_{a0}, and the voltage drop across the load, V_{R0}. The voltages may alternatively be found by the equations

$$V_{R0} = I_{a0}R_L$$

and (14-43)

$$V_{a0} = E_a - V_{R0}$$

Next we find the quiescent anode dissipation, P_{a0}, and check to see if it exceeds the respective

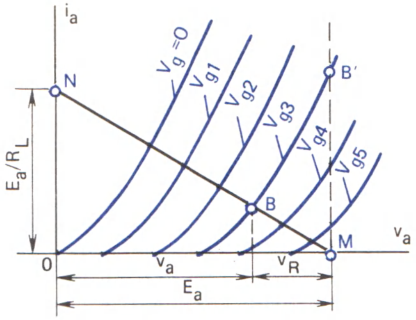


Fig. 14-13

Construction of the load line

absolute maximum rating:

$$P_{a0} = I_{a0}V_{a0} \leqslant P_{a\,max} \tag{14-44}$$

The total power delivered by the anode supply source is

$$P_0 = E_a I_{a0}$$

and the d. c. power dissipated in the load is

$$P_{R0} = I_{a0}V_{R0}$$

or (14-45)

$$P_{R0} = P_0 - P_{a0}$$

The time axis of the alternating grid voltage is shown as an extension to the curve for $v_g = E_g$. As an example, we have taken $V_{mg} = |E_g|$. The peak values of the positive and negative half-cycles of grid voltage correspond to the positive and negative peaks of grid voltage (in our case, to zero and V_{g5}), which mark the end points, A and B, of the *operating region*. Since in the coordinate axes adopted there is no grid voltage axis drawn to a linear scale, this voltage may appear distorted in waveshape, that is, with its positive and negative peaks unequal.

Points A and B mark the positive and negative peak values of the pulsating anode current, $i_{a\,max}$ and $i_{a\,min}$. The waveform of anode current is plotted on the right of the figure.

For distortionless amplification

$$I'_{ma} = I''_{ma} = I_{ma}$$

and (14-46)

$$I_{a\,av} = I_{a0}$$

If, on the other hand, the tube is operating within the nonlinear regions of the characteristic, the positive half-cycle is amplified more than

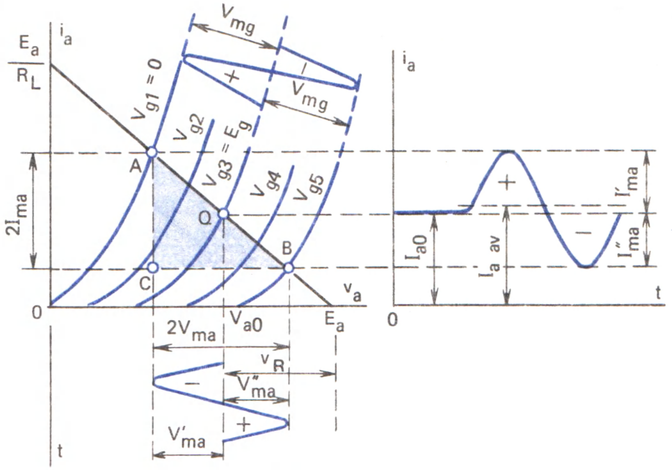


Fig. 14-14

Graphical analysis of triode operation under dynamic conditions (at load)

the negative half-cycle:

$$I'_{ma} > I''_{ma} \qquad (14\text{-}47)$$

Then the amplitude of the useful fundamental is

$$I_{ma} \approx (I'_{ma} + I''_{ma})/2 \qquad (14\text{-}48)$$

or

$$I_{ma} \approx (i_{a\,max} - i_{a\,min})/2 \qquad (14\text{-}49)$$

The amplitude of the second harmonic is

$$I_{ma2} \approx (I'_{ma} - I''_{ma})/4 \qquad (14\text{-}50)$$

or

$$I_{ma2} \approx (i_{a\,max} + i_{a\,min} - 2I_{a0})/4 \qquad (14\text{-}51)$$

The *harmonic ratio* can approximately be found by considering only the second harmonic

$$k_h = I_{ma2}/I_{ma} \qquad (14\text{-}52)$$

Since the positive half-cycle is greater in magnitude than the negative half-cycle, the direct component of anode current, $I_{a\,av}$, exceeds the quiescent current I_{a0}. The incremental change in direct anode current, ΔI_a, is numerically equal to the amplitude of the second harmonic. Hence,

$$I_{a\,av} = I_{a0} + \Delta I_a = (i_{a\,max} + i_{a\,min} + 2I_{a0})/4 \qquad (14\text{-}53)$$

The change in the direct anode current secondary to a transition from the quiescent (no-signal) state to the 'with-signal' mode is an indication of nonlinear distortion. No nonlinear distortion exists when the milliammeter measuring the direct component of anode current gives the same reading in the presence and the absence of an alternating (signal) voltage at the grid. Conversely, nonlinear distortion does exist if the milliammeter gives an increased reading when an alternating (signal) current is applied to the grid.

After I_{ma} and $I_{a\,av}$ have been found, the following quantities can be determined:

- the amplified (output) voltage

$$V_{m\,out} = V_{mR} = V_{ma} = I_{ma}R_L \qquad (14\text{-}54)$$

- the stage voltage gain

$$k_V = V_{ma}/V_{mg} \qquad (14\text{-}55)$$

- the useful (output) power

$$P_{out} = I_{ma}V_{ma}/2 \qquad (14\text{-}56)$$

- the power fed by the anode supply source

$$P_{0\,av} = I_{a\,av}E_a \qquad (14\text{-}57)$$

- the anode-circuit efficiency of the stage

$$\eta = P_{out}/P_{0\,av} \qquad (14\text{-}58)$$

Referring to the anode voltage waveform located below the characteristic curves in Fig. 14-14, it is seen that there is a phase difference of 180° between the alternating grid and anode voltages. A positive half-cycle of grid voltage corresponds to a negative half-cycle of anode voltage, and vice versa. This plot also shows variations in the load voltage, v_R, as counted from the vertical line representing $v_a = E_a$. It is seen that v_a varies in anti-phase with v_R.

Owing to the tube nonlinearity, the positive and the negative half-cycles of the alternating anode voltage differ in amplitude, $V'_{ma} \neq V''_{ma}$. If we know current harmonics, we can multiply them by R_L and obtain anode voltage harmonics.

The useful power is represented by a right triangle, ABC, whose hypotenuse AB bounds the operating region of the stage. When drawn to appropriate scales, its sides give twice the peak values of anode current, I_{ma}, and of anode voltage, V_{ma}. In consequence, the area of the triangle is equal to four times the useful power.

If the load placed in the anode circuit of a tube is a resonant tank circuit or a transformer, the stage curves, including its load line, should be constructed in a different way, that is, as for transistor stages using resonant circuits and transformer coupling (see Sec. 6-1).

14-6 The Vacuum Triode as an Oscillator

A vacuum triode can be used as an oscillator. The circuit of an elementary sinewave oscillator using inductive feedback is shown in Fig. 14-15. It is convenient to treat this oscillator as an amplifier of its own oscillations originally produced in its resonant LC-circuit.

When the anode supply is turned on, free oscillations are produced in the LC-circuit. The alternating voltage existing across the resonant circuit is coupled back to the grid over a feedback coil, L_{fb}, and is amplified by the tube. If positive feedback is used, an amplified voltage is produced across the LC-circuit and sustains its oscillations. For these oscillations to be undamped (that is, for the oscillations to be self-sustaining), two conditions must be satisfied. Firstly, the feedback coil must be connected so that there is a phase difference of 180° between the alternating voltages at the anode and the grid. Secondly, the feedback factor defined as the ratio between the alternating grid and tank-circuit voltages, k_{fb}, must be not less than the reciprocal of the stage voltage gain k_V:

$$k_{fb} \geqslant 1/k_V \tag{14-59}$$

On substituting for k_V from Eq. (14-41), we get

$$k_{fb} \geqslant (R_L + r_a)/\mu R_L = 1/\mu + 1/g_m R_L \tag{14-60}$$

where R_L is the load resistance at resonance.

As is seen, the increase in k_V, μ, g_m and R_L calls for a smaller value of k_{fb} needed for oscillations to sustain themselves.

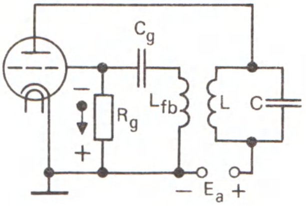


Fig. 14-15

Triode oscillator using inductive feedback

The components R_g and C_g are included to derive a self-bias voltage from the grid current. This grid bias exists only when the stage is generating oscillations. As long as no oscillations are produced, there is no grid current flowing, and the self-bias voltage is nil. When an alternating voltage is applied to the grid, its positive half-cycles give rise to a pulsating grid current. Its direct component produces across R_g a voltage drop which acts as the bias voltage. The capacitor C_g serves to reduce the ripple in this voltage. This bias scheme enhances the stability of oscillator operation.

14-7 The Interelectrode Capacitances of a Vacuum Triode

The performance of a vacuum triode is markedly affected by its interelectrode capacitances whose magnitude depends on the design and operating conditions of the tube. Their effect is stronger at higher frequencies.

There are three interelectrode capacitances associated with the vacuum triode. In circuit diagrams they are shown by capacitor symbols drawn in dashed lines as in Fig. 14-16*a*. The grid-to-cathode capacitance, C_{gk}, is sometimes called the input capacitance; the anode-to-cathode capacitance, C_{ak}, is the output capacitance; and the anode-to-grid capacitance, C_{ag}, is the transfer capacitance. For low- and medium-power tubes, these capacitances usually run into units of picofarads. The figures quoted in tube manuals include both the interelectrode and the interlead capacitances. There is always a spread in magnitude between the interelectrode capacitances of individual tubes of the same type.

Let us see how these capacitances affect the performance of the vacuum triode.

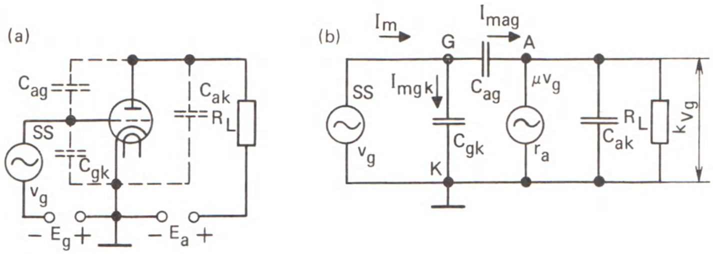


Fig. 14-16

Triode amplifier stage: (*a*) interelectrode capacitances and (*b*) a complete equivalent circuit incorporating all interelectrode capacitances

With a sufficiently high grid bias, it would seem that the grid should draw no current. Owing to the input capacitance C_{gk}, however, a capacitive current exists in the grid circuit. Thus, the input capacitance loads the signal source, *SS*. This capacitive current produces a voltage drop across the internal resistance R_{ss} of the signal source. As a result, there is a decrease in the useful voltage across the signal-source terminals or, which is the same, across the tube input, and also in the alternating anode current, the amplified alternating anode voltage, and the useful output power. As we move to higher frequency, the reactance of the input capacitance decreases, the grid draws a progressively larger current, and the voltage drop across R_{ss} increases.

This effect is not noticeable at low frequencies. At high frequencies, however, it may appreciably degrade the efficiency of an amplifier stage. The effect of C_{gk} may only be neglected at frequencies where the reactance $1/\omega C_{gk}$ is many times R_{ss}. Let, for example, $R_{ss} = 100$ kΩ and $C_{gk} = 10$ pF. Then at a frequency of 500 Hz, $1/\omega C_{gk} = 32$ MΩ. This reactance is equivalent in its effect to an open circuit. If, however, we raise the frequency to 5 MHz, that is, by a factor of 10^4, the reactance presented by the input capacitance will be 3.2 kΩ. It wll heavily load the signal source, and its voltage will be brought down drastically.

The anode-to-cathode, or output, capacitance C_{ak} shunts the stage load. This is clearly seen from reference to the stage equivalent circuit in Fig. 14-16*b* where all of the interelectrode capacitances are taken care of. The tube is replaced by an equivalent constant-voltage generator (the Thévenin equivalent), which generates an emf equal to μV_g and has an internal resistance R_i. It operates into a load z_L made up of a parallel connection of R_L and C_{ak}. At low frequencies, $1/\omega C_{ak}$ is many times R_L, so it is legitimate to think of the load impedance as consisting solely of R_L. At high frequencies, however, $1/\omega C_{ak}$ may be of the same order of magnitude as R_L or even smaller. The load impedance z_L will be less than R_L, and this will reduce the stage gain. The greater the value of R_L, the stronger the shunting effect of the tube's output capacitance, and the lower the frequencies at which this effect is felt.

At high frequencies, C_{ak} produces a further phase shift in the output voltage. This is of no consequence in a. f. amplifiers, but it is entirely intolerable for TV and some other signals.

In stages loaded by a tank circuit (such as in r. f. amplifiers and oscillators), C_{ak} is part of the tank circuit and is lumped with the tank-circuit capacitance. Therefore, in designing such a stage, the actual capacitor is chosen to be smaller by an amount equal to C_{ak}. At very high frequencies, it may prove that the tank-circuit capacitance must be smaller than C_{ak}. Such a tank cannot be built. If there is a resonant circuit in the grid lead, the input capacitance of the tube is lumped with the grid tank-circuit capacitance.

Because of the spread in interelectrode capacitances among the tubes of the same type, a tube change may upset the tuning of the tank circuit.

The ability of a tube to be used over a wide range of frequencies with an equal performance is usually stated in terms of its *bandwidth factor*, γ, defined as the ratio of the tube's transconductance to the sum of its input and output capacitances:

$$\gamma = g_m/(C_{gk} + C_{ak}) \qquad (14\text{-}61)$$

The greater the value of γ, the better the tube. Good tubes have a bandwidth factor of 1-3 mA V^{-1} pF^{-1}. The value of γ can mostly be improved by raising the tube's transconductance because its interelectrode capacitances cannot be reduced by a substantial amount.

The anode-to-grid, or transfer, capacitance C_{ag} is the worst offender of all. As follows from the circuit of Fig. 14-16*b*, the capacitive current I_m which loads the signal source *SS* is equal to the sum of the capacitive currents I_{mgk} and I_{mag} flowing through C_{gk} and C_{ag}, respectively:

$$I_m \approx I_{mgk} + I_{mag} \qquad (14\text{-}62)$$

This is an approximate equality because I_{mag} is not purely capacitive owing to the presence of the resistances R_L and R_i. The currents must be added vectorially rather than arithmetically.

By Ohm's law,

$$I_{mgk} = V_{mg}\omega C_{gk}$$

and (14-63)

$$I_{mag} = V_{mag}\omega C_{ag}$$

where V_{mag} is the peak value (amplitude) of the anode-to-grid voltage.

When a tube is operating into a resistive load, the alternating grid and anode voltages, V_{mg} and V_{ma}, are 180° out of phase with each other, and V_{mag} is equal to the sum of these two voltages

$$V_{mag} = V_{mg} - (-V_{ma}) = V_{mg} + V_{ma} \qquad (14\text{-}64)$$

On taking V_{mg} out of the brackets, we get

$$V_{mag} = V_{mg}(1 + V_{ma}/V_{mg})$$
$$= V_{mg}(1 + k_V) \qquad (14\text{-}65)$$

Hence,

$$I_m = V_{mg}\omega C_{gk} + V_{mg}\omega C_{ag}(1 + k_V)$$
$$= V_{mg}\omega[C_{gk} + C_{ag}(1 + k_V)] \qquad (14\text{-}66)$$

The expression in the square brackets gives the input capacitance of an amplifier stage built around a vacuum triode in the 'with-signal' state:

$$C_{\text{in s}} = C_{gk} + C_{ag}(1 + k_V) \qquad (14\text{-}67)$$

At no-load (in the 'no-signal' condition), $k_V = 0$ and the input capacitance of an amplifier stage is

$$C_{\text{in}} = C_{gk} + C_{ag} \qquad (14\text{-}68)$$

Thus, the input capacitance of an amplifier stage at load (the 'with-signal' condition) is many times the figure at no-load (the 'no-signal' condition). For example, if $C_{gk} = 5$ pF, $C_{ag} = 3$ pF and $k_V = 40$, then at no-load (the 'no-signal' condition):

$$C_{\text{in}} = 5 + 3 = 8 \text{ pF}$$

and at load (the 'with-signal' condition):

$$C_{\text{in s}} = 5 + 3(1 + 40) = 128 \text{ pF}$$

Thus, there is a 16-fold increase in capacitance.

A further detrimental effect of C_{ag} consists in that it transfers some of the signal current into the anode circuit. Hence, its name 'transfer capacitance'. In some applications the signal source is operating continuously while the tube is driven to cutoff at regular intervals during which there must be no alternating anode current flowing. Due to C_{ag}, however, some current finds its way from the signal source to R_L even when the tube is at cutoff.

A third effect of the transfer capacitance is especially troublesome – it consists in feedback from the anode to the grid circuit via C_{ag}. The amplified oscillations find their way via C_{ag} from the anode circuit back to the grid circuit. It is seen from Fig. 14-16*b* that the alternating current supplied by the equivalent generator which replaces the tube goes not only to R_L, but also via C_{ag} into the grid circuit. This current produces across the grid-cathode space a feedback voltage which is added to that supplied by the signal source *SS*.

It may be said that across the anode-cathode space the output voltage is applied to a divider consisting of C_{ag} and the grid-cathode space. Some of the voltage existing across this portion of the circuit is a feedback voltage. Its magnitude is different, depending on the relative magnitudes of the resistances presented by these portions of the circuit. At higher frequencies, C_{ag} decreases and the feedback voltage is increased. In the case of positive feedback, this may lead to the generation of spurious oscillations. When, however, a stage intended to operate as an amplifier turns into an oscillator, this is an indication that the operation of the stage has completely been dislocated. The generation of oscillations due to feedback via C_{ag} comes about readily in stages whose anode and grid leads contain resonant circuits. This is the reason why instead of triodes, r. f. amplifiers use tetrodes and pentodes in which the detrimental effect of the transfer capacitance is suppressed (see Chap. 15).

14-8 Common-Grid and Common-Anode Stages

So far, we have dealt with the common-cathode amplifier stage widely used in many applications. Sometimes, however, resort is made to *common-grid* and *common-anode stages.*

The circuit of a common-grid amplifier stage is shown in Fig. 14-17. It has several special features that deserve mention. There is no current amplification ($k_I = 1$), and so $k_P = k_V$. A disadvantage of this circuit is that it has a low input resistance because the input current is the cathode current. R_{in} is approximately equal to $1/g_m$. For example, if $g_m = 5$ mA V^{-1}, then $R_{\text{in}} = 1/5 = 0.2$ kΩ. The control grid doubles as a screen grid. Because of this, C_{ak} which acts as a transfer capacitance is very small in magnitude. For this reason, the common-grid stage is especially suitable for operation at microwave frequencies.

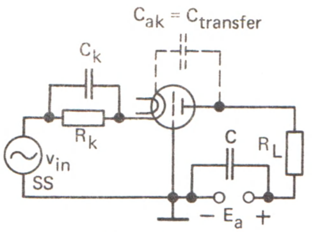


Fig. 14-17

Common-grid amplifier stage

The circuit of a common-anode stage is shown in Fig. 14-18. Alternatively, it is known as the *cathode follower* because its load R_L is placed in the cathode lead and the output voltage follows the input voltage very closely in both magnitude and phase. This circuit does not supply voltage amplification ($k_V \approx 1$) but current amplification is considerable, so $k_P \approx k_I$. An advantage of this circuit is its low input capacitance. The common-anode stage has a very stable gain and produces negligible distortion in the amplified signal. These advantages are due to a large amount of negative feedback ($k_{\text{fb}} = 1$). All of the output voltage is passed back to the input. Cathode followers are especially often used in pulse circuitry because they do not practically distort the waveform of the amplified pulses.

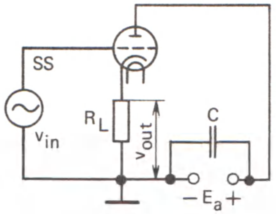


Fig. 14-18

Common-anode amplifier stage (cathode follower)

14-9 Limitations of the Vacuum Triode

The vacuum triode is prone to three major limitations. Firstly, it cannot combine a high amplification factor with a large cutoff voltage. If, in an attempt to secure a high value of μ, we build a triode with a very short-pitch grid, it will be driven to cutoff by a very small negative grid voltage. For example, if μ = 1000, then at $V_a = 250$ V the grid cutoff voltage will be as small as

$$V_{\text{g co}} = -V_a/\mu = -250/1000 = -0.25 \text{ V} \qquad (14\text{-}69)$$

In this case, nearly all of the anode-grid characteristic will lie within the region of positive grid voltages. Such a tube will obviously draw a heavy grid current in operation. If we wished to shift the anode-grid characteristic of this tube to the left (that is, into the region of negative grid voltages), we would have to make its anode voltage prohibitively high. For example, if we wished that at μ = 1000 the cutoff voltage were − 5 V, we would have to set the anode voltage at

$$V_a = -\mu V_{\text{g co}} = -1000 \times (-5) = 5000 \text{ V}$$

This is the reason why vacuum triodes are built with an amplification factor of not over 100. For a distortion-free amplification of strong a.f. signals, vacuum triodes should have a low amplification factor so that their anode-grid curves could lie within the region of negative grid voltages.

A second limitation is that the a.c. anode resistance, r_a, of a vacuum triode is relatively small. In r.f. amplifiers containing tank circuits, the a.c. anode resistance of the tube shunts the anode tank (see Fig. 14-12), thereby impairing its resonant properties and Q-factor. The smaller the value of r_a, the stronger it shunts the tank circuit and the greater it impairs the performance of the tank circuit. This is the reason why tubes for r. f. amplifiers have to have a high a. c. anode resistance r_a.

The third limitation of the vacuum triode is its relatively high transfer capacitance C_{ag}. Its detrimental effect has been discussed earlier.

The limitations of the vacuum triode can be eliminated almost completely by adding a second grid. This will be discussed in more detail in Chap. 15.

14-10 Basic Types of Receiving-Amplifying Vacuum Triodes

Most often, vacuum triodes are built to handle small amounts of power and are used for a. f. amplification and detection. Many vacuum triodes are used as oscillators and r. f. amplifiers set up in configurations capable of minimizing the effect of the transfer capacitance (an example is the common-grid circuit). Wide use is made of dual triodes. There is a special group of what are known as *feed-through triodes* intended mainly for use in electronic voltage stabilizers. Typically, they have a low a. c. anode resistance, a small amplification factor, but a high transconductance. Electronic voltage stabilizers also use H. V. feed-through triodes with a very small transconductance and very high values of μ and r_a.

A good deal of time and effort has been put into the improvement of tube's transconductance. This parameter is pivotal in applications where it is essential to produce an undistorted amplified signal, such as TV, radar, and automatic process control. One way to improve the transconductance of a tube is to reduce its grid-to-cathode spacing. The transconductance is inversely proportional to the square of this distance. The point is that the potential barrier is located very closely to the cathode. For an effective control of the electron stream, the grid must be placed as near the potential barrier as practicable. Improvements in tube manufacture have made it possible to build tubes in which the grid-to-cathode spacing is a few tens of micrometres. The grid is made of wires 7-10 μm in diameter and spaced very closely apart so as to minimize the island effect. The transconductance of these tubes is several tens of milliamperes per volt.

Chapter Fifteen

Tetrodes, Pentodes and Miscellaneous Tubes

15-1 The Vacuum Tetrode

A *vacuum tetrode* is a tube which has a *second*, or *screen*, grid. It is a mesh of fine wire placed between the control grid and the anode. The screen grid serves a three-fold purpose: it improves the amplification factor μ of the tube, increases its a. c. anode resistance r_a, and minimizes its transfer capacitance C_{ag}.

The quantities associated with the first, or control, grid, have the subscript '$g1$', and the quantities related to the second, or screen, grid, have the subscript '$g2$'.

If the screen grid is connected to the cathode, it shields both the cathode and the control grid against the action of the anode. The screen grid intercepts the greater proportion of the electric field set up by the anode. Only a tiny fraction of the electric-field flux originating at the anode is able to pass through the screen grid. How much the screen grid weakens the electric field of the anode is expressed in terms of the screen-grid penetration factor D_2.

The electric flux breaking through the screen grid is then intercepted by the control grid so that only a fraction of the original flux goes through. The extent to which the control grid attenuates the electric field depends on the control-grid penetration factor D_1. In the final analysis, only an extremely tiny proportion of the original electric flux produced by the anode is able to pass through the two grids and to reach the potential barrier at the cathode. Quantitatively, it is stated in terms of the *tetrode penetration factor D* defined as the product of the screen-grid and control-grid penetrations:

$$D = D_1 D_2 \qquad (15\text{-}1)$$

The tetrode penetration factor tells how much weaker the anode voltage is in its effect on the cathode current than the control-grid voltage. For example, if $D = 0.01$ this means that a change of 1 volt in anode voltage produces on the cathode current an effect which is one-hundredth of the effect produced by the same change in grid voltage. In other words, the effect produced on the cathode current by an anode voltage of 1 volt is equivalent to the effect produced by a control-grid voltage of 0.01 volt.

Approximately, the penetration factor is the reciprocal of a tube's amplification factor (see p. 168):

$$\mu \approx 1/D = 1/D_1 D_2 \qquad (15\text{-}2)$$

If the screen grid lets through 2% of the total electric flux originating at the anode and the control grid lets through a mere 10% of the remaining flux, the cathode will see only 0.2% of the original electric flux. All in all, the action of the anode on the potential barrier at the cathode is weakened by a factor of 500. That is, the amplification factor of the tube is about 500.

The amplification factor of vacuum tetrodes may be as high as several hundred, and the a. c. anode resistance may be several hundred kilohms.

To sum up, using two grids with a moderately short pitch, we can assure a high amplification factor and a large a. c. anode resistance. Also, by applying an appreciable positive voltage to the screen grid, we can shift the anode-grid characteristics into the region of negative grid voltages. How this comes about will be clear from the discussion that follows.

To begin with, we will examine the equivalent voltage of a tetrode. It is 'equivalent' in the sense that its effect is equal to that produced by the joint action of the anode, screen-grid and control-grid voltages. This equivalent (or *composite*) voltage, V_{eq}, is assumed to be applied across an equivalent diode taking the place of the control grid. By analogy with the vacuum triode we may write

$$V_{eq} \approx V_{g1} + D_1 V_{g2} + D_1 D_2 V_a \qquad (15\text{-}3)$$

Equation (15-3) shows that the action of the screen grid is solely weakened by the control grid (V_{g2} is multiplied by D_1), while the action of the anode is weakened by both grids (V_a is multiplied by $D_1 D_2$).

Now we are able to write the three-halves power law for the vacuum tetrode:

$$i_k = k V_{eq}^{3/2} \qquad (15\text{-}4)$$

where k, or the perveance, is a function of the electrode geometry as is the case with the vacuum triode.

The cathode current in a tetrode is the sum of the anode, screen-grid and control-grid currents:

$$i_k = i_a + i_{g2} + i_{g1} \qquad (15\text{-}5)$$

When the control grid is held at a negative potential, $i_{g1} = 0$ and

$$i_k = i_a + i_{g2} \qquad (15\text{-}6)$$

The screen grid is held at a positive direct voltage which is equal to 20-50% of anode voltage. This voltage brings down the potential barrier at the cathode, so that electrons could travel towards the anode. Because of the two intervening grids, the anode has but a very weak influence on the potential barrier near the cathode. When the screen-grid voltage is zero and the control grid is held at a negative potential, a retarding field appears in the control grid-cathode space. The equivalent voltage is thus negative and the potential barrier at the cathode is so high that electrons cannot climb over it. In consequence, when $V_{g2} = 0$, the tube is at cutoff, and its anode current is zero. Let, for example, $V_{g1} = -3$ V, $V_{g2} = 0$, $V_a = 300$ V, and $D = 0.002$. Then,

$$V_{eq} = -3 + 0.002 \times 300 = -2.4 \text{ V}$$

The screen-grid current i_{g2} is constituted by electrons which are able to reach this grid. Under normal operating conditions when the anode voltage is higher than the screen-grid voltage, i_{g2} is a small fraction of anode current because the majority of electrons move through the screen grid at a high velocity.

The term $D_1 D_2 V_a$ in Eq. (15-3) may be neglected because $D_1 D_2 \ll 1$. Therefore,

$$V_{eq} \approx V_{g1} + D_1 V_{g2} \qquad (15\text{-}7)$$

For the tube to be driven to cutoff, the equivalent (composite) voltage must be zero, $V_{eq} = 0$. Then, $i_k = 0$. The voltage that must exist at the grid of a tetrode in order to drive it to cutoff can be found from Eq. (15-7):

$$V_{g1\,co} \approx -D_1 V_{g2} \qquad (15\text{-}8)$$

Because the wires of the control grid are spaced widely apart and V_{g2} is rather high, it follows that the cutoff voltage must be high – this means that the anode-grid characteristics of the tetrode are shifted into the region of negative grid voltages. For example, if $D_1 = 0.1$, $D_2 = 0.02$ and $V_a = 250$ V, then at $V_{g2} = 100$ V,

$$V_{g1\,co} = -0.1 \times 100 = -10 \text{ V}$$

In view of the effect produced by the anode, we get

$$V_{g1\,co} = -0.1 \times 100 - 0.002 \times 250 = -10.5 \text{ V}$$

In this case, an appreciable portion of the anode-grid characteristic in the range 0 to − 10 V lies within the region of negative grid voltages. For a triode with $D = 0.002$ and $V_a = 250$ V, we will have $V_{g\,co} = -0.5$ V which means that the characteristic of the triode will lie in the region of positive grid voltages.

How the screen grid serves to reduce C_{ag1} can be elucidated by reference to the simplified equivalent circuit of Fig. 15-1. The supply voltage sources are omitted because the equivalent circuit has solely been drawn up for the capacitive alternating current. If there were no screen grid, the grid and anode circuits would be coupled by the transfer capacitance C_{ag1}. In the presence of the screen grid G_2 connected to the cathode, the capacitive current produced by the signal source may take any one of two paths. One runs from G_2 via the wire connecting G_2 to the cathode and back to the signal source. The other runs from G_2 via the screen grid-to-anode capacitance and the load resistance R_L back to the signal source. The second path presents an opposition which is many times that presented by the first. For this reason, nearly all of the capacitive current i_C chooses the first path. In this way, the capacitive coupling between the grid and anode circuits is all but eliminated.

The reduction in the transfer capacitance may be interpreted in a different way. The charges at the anode and control grid interact via an electric field. In a tetrode, the screen grid intercepts the greater proportion of the electric flux produced by the charge on the anode. The field that couples the charges on the anode and control grid is greatly reduced, and this means that the capacitance between these two electrodes is drastically reduced. For example, if the screen grid lets through a mere 2% of the flux originating at the anode, the interaction between the charges on the anode and control grid will be weakened by a factor of 50, and C_{ag1} will be reduced likewise by a factor of 50.

The transfer capacitance of a tetrode is reduced in about the same proportion as the tube's amplification factor is increased. When the screen grid is wound with a short pitch, it reduces the transfer capacitance to a greater extent than when it is wound with its wires spaced wider apart.

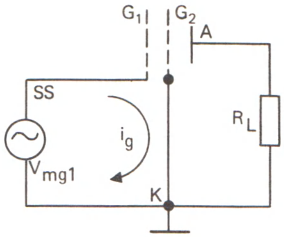


Fig. 15-1

Equivalent circuit illustrating reduction in transfer capacitance by the screen grid

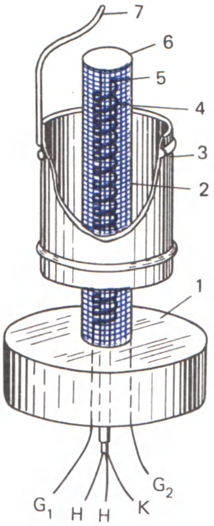


Fig. 15-2

Tetrode: (*1*) shield; (*2*) screen grid; (*3*) anode; (*4*) control grid; (*5*) cathode; (*6*) shield; (*7*) anode lead

Some of the transfer capacitance still does remain because a fraction of the original electric flux originating at the anode reaches the control grid in a round-about way rather than directly through the screen grid. It can be eliminated or, at least, minimized by placing a solid metal shield to intercept the remaining flux. A tetrode built along such lines, with its anode cut out in part, is shown in Fig. 15-2.

A transfer capacitance also exists between the anode leads and the control-grid leads. One way to reduce it is to space the leads wider apart and to put in another shield between them. The anode leads can be carried to the top of the tube envelope and the control-grid leads to the tube base (or the other way around). The anode circuit can additionally be shielded from the grid circuit outside the tube. The control-grid lead can be enclosed in a shielding sleeve, and the elements of the two circuits (anode and control-grid) can further be isolated from one another by shielding.

15-2 Secondary Electron Emission in the Vacuum Tetrode

A major limitation of the vacuum tetrode is that it is strongly influenced by the secondary-electron interchange between the anode and the screen grid, known as the *dynatron effect* in the Soviet literature of the subject. The electrons striking the anode knock secondary electrons out of it. This secondary electron emission from the anode exists in all tubes, but it does not give rise to any detrimental consequences in the diode or the triode. In them, the secondary electrons leaving the anode fall back because the anode is more positive, so there is no current flowing due to secondary electrons.

In a tetrode, secondary electron emission from the anode does not manifest itself as long as the screen grid is held at a lower potential than is the anode because the secondary electrons return to the anode. If in the dynamic operation, when a tetrode is operating at load (in the 'with-signal' condition), an increase in anode current may cause the anode to become during certain time intervals less positive than the screen grid to which a direct voltage is applied. Then the secondary electrons emitted by the anode will be attracted by the screen grid, giving rise to a secondary-electron current opposite in direction to the primary-electron current. The total anode current is reduced and the screen-grid current is increased. This is the consequence of what we have referred to as the dynatron effect. Figure 15-3 shows the electron streams corresponding to the current i_{aI} constituted by the primary electrons reaching the anode, the screen-grid current i_{g2I} constituted by primary electrons, and the secondary-electron current i_{II} travelling from the anode to the screen grid. The resultant anode and screen-grid currents are given by

$$i_a = i_{aI} - i_{II}$$

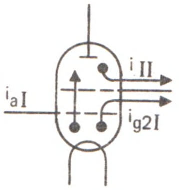


Fig. 15-3

Currents in a tetrode in the presence of secondary emission

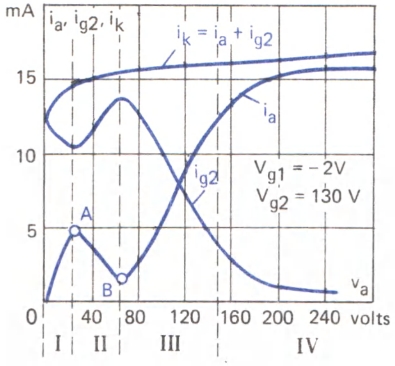


Fig. 15-4

Anode, screen-grid and cathode characteristics of a tetrode

and (15-9)

$$i_{g2} = i_{g2I} + i_{II}$$

If the secondary-emission ratio is more than unity, i_{II} may exceed i_{aI}, and the anode current will become negative.

The dynatron effect ought not to be construed as to be due solely to secondary electron emission. Secondary electron emission is just a necessary, but not a sufficient condition for the dynatron effect to take place. Another necessary condition is that the anode must be at a lower potential than the screen grid. If secondary emission occurs but the second condition is not satisfied, no dynatron effect will take place.

The dynatron effect markedly affects the anode characteristics of tetrodes. Figure 15-4 shows plots of a tetrode's anode current, screen-grid current and cathode current as functions of anode voltage, with the control and screen grids held at constant potentials.

As is seen, four regions may be noted in the anode characteristics of a tetrode. Region *I* corresponds to low anode voltages (10-20 V).

No secondary electrons are emitted from the anode because the primary electrons are travelling at a low velocity. As the anode voltage is raised, there is an abrupt rise in anode current and a reduction in the screen-grid current (the fall-back mode). The anode voltage has a negligible effect on the cathode current because the electric field set up by the anode acts on the potential barrier at the cathode through the two grids. Therefore, the cathode current changes only slightly.

When the anode voltage rises above 10-20 V, secondary emission takes place, and the dynatron effect occurs. As the anode voltage is raised still more, a progressively greater number of secondary electrons are emitted, the anode current falls, and the screen-grid current rises (region *II*). With a heavy secondary emission, point *B* may be located below the axis of abscissas. The dynatron effect ceases when $v_a = v_{g2}$, that is, it occurs within regions *II* and *III* of the anode characteristic.

When the anode voltage exceeds the screen-grid voltage (region *IV*), there occurs a small rise in anode current and a small fall in the screen-grid current. Secondary emission does exist within this region, but all of the secondary electrons return to the anode, and no dynatron effect is observed. Tetrodes are usually operated within region *IV* where large changes in anode voltage bring about negligible changes in anode current. This means that within region *IV* a tetrode has high values of μ and r_a. It ought not to be construed that the quietly sloping curves within this region is an indication of saturation (or a temperature-limited condition). Rather, the current is limited by the potential barrier that exists at the cathode.

The characteristic has a down-sloping '*kink*', *AB*, within which the tetrode has a negative a. c. plate resistance because a positive change in anode voltage, Δv_a, gives rise to a negative change in anode current, Δi_a:

$$r_{a\,AB} = \Delta v_a / \Delta i_a < 0 \qquad (15\text{-}10)$$

A negative-resistance device can operate as an oscillator. This is also true of the tetrode – it is able to generate an alternating current within the down-sloping kink of its characteristic.

The dynatron effect is an undesirable feature of tetrodes. The strong nonlinearity of its anode characteristics gives rise to heavy nonlinear distortion in the amplified signal. Nor is it an advantage when the screen-grid current exceeds the useful anode current. Also, the tube may generate unwanted, stray oscillations. One way to avoid the dynatron effect is to hold the screen grid at a potential which is always lower than that at the anode (it is usually 20-50% of anode voltage).

15-3 The Pentode

The *pentode* came as a further step in the development of vacuum tubes, following the advent of the tetrode. It has the advantages of the tetrode and is free from the dynatron effect.

The pentode, as its name implies, is a five-element tube. It is like a screen-grid tetrode with one more grid added between the screen grid and the anode. The additional grid is called the *suppressor grid* for the reason that it suppresses secondary electron emission. Sometimes it is referred to as the *third grid*, G_3, and the quantities associated with it are usually given the subscript '$g3$'.

As a rule, the suppressor grid is connected to the cathode which means that it is held at zero potential with respect to the cathode and a negative potential with respect to the anode. In some circuits, a small positive or negative voltage may be applied to the suppressor grid. In such a case, however, its potential is below that of the anode, and the dynatron effect persists. In our further discussion it will be assumed that $v_{g3} = 0$. In many pentodes, the suppressor grid is connected to the cathode inside the envelope. If so, the suppressor-grid voltage is always zero. The action of the suppressor grid consists in that the electric field set up between the grid and anode first retards, then brings to a stop, and finally drives back to the anode all the secondary electrons knocked out of it. The electrons are unable to reach the screen grid, and the dynatron effect is thus suppressed.

The electrons travelling from the cathode see a retarding field between the screen and suppressor grids, and it would seem that such a field should cause a fall in anode current. Actually, the electrons brought up to a high velocity by the screen grid are able to reach the suppressor grid without losing their speed because there is a positive rather than a negative potential difference between the suppressor-grid wires. Figure 15-5 shows the potential profile in the screen grid-anode space of a pentode. As an example, it is assumed that $V_{g2} = 200$ V, $V_{g3} = 0$, and $V_a = 100$ V. The suppressor-grid wires are at zero potential, and the potential between the wires is positive but lower than it is at the anode (+ 50 V). The potential distribution curve, *1*, along line *BC* which passes through a suppressor-grid wire, and the potential distribution curve, *2*, along line *DE* which passes midway between two adjacent suppressor-grid wires, show that there is a second potential barrier over which the secondary electrons knocked out of the anode cannot climb.

Pentodes have a higher amplification factor which runs sometimes into several thousand. This is because the suppressor grid acts as an additional screen grid. The a. c. anode resistance is likewise greater, being sometimes of the order of several megohms. The transfer capacitance is smaller than it is in tetrodes. The transconductance of pentodes is about the same as in triodes and tetrodes, that is, anywhere between 1 and 50 mA V^{-1}.

The equivalent (composite) voltage of a pentode is given by

$$v_{eq} \approx v_{g1} + D_1 v_{g2} + D_1 D_2 v_{g3} + D_1 D_2 D_3 v_a \tag{15-11}$$

Hence, the penetration factor of a pentode is

$$D = D_1 D_2 D_3 \tag{15-12}$$

and is very small.

Because D is small and the third term is either zero if $v_{g3} = 0$, or very small because $D_1 D_2 \ll 1$, it follows that the equivalent and cutoff voltages are defined similarly to the same quantities for tetrodes:

$$v_{eq} \approx v_{g1} + D_1 v_{g2}$$

and (15-13)

$$v_{g1\,co} \approx -D_1 v_{g2}$$

The anode-grid characteristics of pentodes are similar to those of tetrodes, that is, they lie in the region of negative grid voltages.

The three-halves power law for a pentode has the form

$$i_k = k v_{eq}^{3/2} \tag{15-14}$$

where the cathode current is

$$i_k = i_a + i_{g1} + i_{g2} + i_{g3} \tag{15-15}$$

When the control grid is held negative, $i_{g1} = 0$. The suppressor-grid current i_{g3} has to be taken into account only when $v_{g3} > 0$. Therefore, the cathode current is in most cases the sum of two currents, as it is in a tetrode:

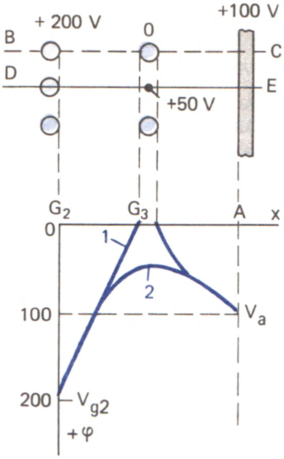


Fig. 15-5

Potential diagram for the anode-screen grid space of a pentode

$$i_k = i_a + i_{g2} \tag{15-16}$$

Sometimes, the suppressor grid is utilized as a second control grid. Also, a pentode can be used to do the job of two tubes. Then, one stage will use the triode section of a pentode (its cathode and the control and screen grids), and another stage will use all of the pentode.

15-4 Current Division in the Pentode

Current division in a pentode may be described in terms of the current division ratio k_c, defined as the ratio between the anode current and the cathode current. If $i_{g1} = 0$ and $i_{g3} = 0$, then

$$k_c = i_a/(i_a + i_{g2}) \tag{15-17}$$

The value of k_c depends on the pentode design and the ratio v_a/v_{g2}. If $v_a = 0$, then $i_a = 0$, and $k_c = 0$. All of the electrons that constitute the cathode current will then be collected by the screen grid. This happens due to the interception of electrons by the screen grid and the return of the electrons that move past the grid but are unable to overcome the retarding field in the screen grid-anode space.

In the fall-back mode, a change in v_a or v_{g2} brings about a change in the height of the second potential barrier, and this leads to a marked

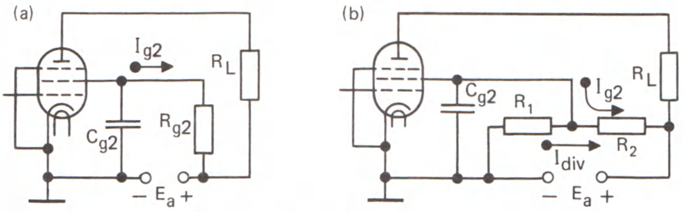


Fig. 15-6

Screen-grid supply circuit: (*a*) via a swamping resistor and (*b*) from a divider

changes in i_a, i_{g2} and the current division ratio k_c. For example, a rise in v_a entails a fall in the potential barrier, and many of the electrons that move past the screen grid do not go back any longer but keep moving on until they are collected by the anode. The anode current rises, the screen-grid current falls, and k_c increases. The second potential barrier is located near the suppressor grid and is subjected to a strong influence from the anode so that its height varies appreciably even when the anode voltage changes only slightly. Therefore in the fall-back mode k_c changes abruptly following a change in the ratio v_a/v_{g2}. In the intercept mode, k_c changes only slightly even when the ratio v_a/v_{g2} is varied by a large amount because all of the electrons moving towards the anode climb over the second potential barrier and any change in the height of this barrier does not affect the current division.

The boundary between the fall-back and intercept modes corresponds to different values of the ratio v_a/v_{g2} for different tubes (it is usually from 0.1 to 0.5), depending on the electrode structure and the voltages at the other grids.

The suppressor-grid voltage strongly affects the current division ratio in the fall-back mode. If the suppressor grid is made more negative, the second potential barrier will increase in height, and k_c will decrease because a progressively smaller number of electrons are able to reach the anode.

15-5 Connection of Tetrodes and Pentodes in Circuits

What sets tetrodes and pentodes apart from other tubes is their screen grid circuit. In pentodes, the screen grid may be held at any potential because the dynatron effect is negligible. In low-power stages, the screen grid is usually held at a low voltage (20-50% of the anode voltage) because there is no need in a large anode current when weak signals are amplified. In higher-power stages the anode current must be greater, and $V_{g2\,0}$ is set at higher values. In power pentodes, the screen-grid voltage may sometimes be set at $V_{g2\,0} = E_a$.

Sometimes, a separate supply source is used for the screen grid of pentodes used in high-power stages. In low-power and multi-tube circuits this scheme is wasteful. However, it offers an advantage in that $V_{g2\,0}$ is held at a constant value. In some cases, $V_{g2\,0}$ is derived as a fraction of the anode supply voltage.

Most often, the screen grid is energized via a voltage-dropping resistor R_{g2} (Fig. 15-6*a*) whose value may range from units to hundreds of kilohms. In this circuit configuration

$$V_{g2\,0} = E_a - I_{g2\,0}R_{g2} \tag{15-18}$$

If we know the screen-grid current, the resistance required to obtain the desired value of $V_{g2\,0}$ will be given by

$$R_{g2} = (E_a - V_{g2\,0})/I_{g2\,0} \tag{15-19}$$

For example, if $E_a = 160$ V, $V_{g2\,0} = 60$ V and $I_{g2\,0} = 0.5$ mA, we will need

$$R_{g2} = (160 - 60)/0.5 = 200\ \text{k}\Omega$$

A limitation of the above arrangement is that $V_{g2\,0}$ changes whenever the operating conditions of the tube are varied. To demonstrate, if we change the filament (heater) voltage, the anode voltage, or the control-grid voltage, this will bring about a change in $I_{g2\,0}$. In consequence, there will also be a change in the voltage drop across R_{g2} and in the screen-grid voltage.

The stability of the screen-grid voltage is greatly improved by the use of a voltage divider made up of two resistors, R_1 and R_2, connected in series as shown in Fig. 15-6*b*. These resistors with a resistance of tens of kilohms pass a current, I_{div}. The voltage produced by this divider current across R_1 is applied to the screen grid. The arrangement using a voltage divider is less economical because I_{div} is wasted as heat in the divider. The greater the value of I_{div} in comparison with $I_{\text{g2 0}}$, the better the stability of $V_{\text{g2 0}}$, but the divider itself dissipates more power.

The values of R_1 and R_2 can be found by the equations

$$R_1 = V_{\text{g2 0}}/I_{\text{div}}$$

and (15-20)

$$R_2 = (E_{\text{a}} - V_{\text{g2 0}})/(I_{\text{g2 0}} + I_{\text{div}})$$

Suppose that we are to design a divider to derive a voltage $V_{\text{g2 0}} = 80$ V from an anode supply source delivering $E_{\text{a}} = 240$ V with $I_{\text{g2 0}} = 1$ mA and I_{div} set at 4 mA. Using the above equations, we find that $R_1 = 80/4 = 20$ kΩ and $R_2 = 160/5 = 32$ kΩ.

As a way of reducing the transfer capacitance, the screen grid is usually connected to the cathode (the common ground) via a sufficiently large capacitor which should present a very small impedance. At high frequencies, the capacitance may be as low as thousands or tens of thousands of picofarads, while at low frequencies the figure should be a few tenths of a microfarad or even greater. Such a capacitor practically short-circuits the screen grid to the cathode for alternating current.

In the absence of such a capacitor, alternating current can be coupled from the control-grid circuit into the anode circuit by the interelectrode capacitances C_{g2g1} and C_{ag2} (Fig. 15-7). When C_{g2} is included in the circuit, the alternating current can flow from the grid circuit via the interelectrode capacitance C_{g2g1} past which it may take any one of two paths, namely: via C_{g2} presenting a very low impedance, or via the interelectrode capacitance C_{ag2} presenting a very high impedance, and then through the load having a high impedance. Obviously, nearly all of the current follows the first path, with only a very negligible fraction dividing into the second path. Thus, the screen grid and the capacitor C_{g2} eliminate capacitive coupling between the anode and grid circuits.

The capacitor C_{g2} does one more job. In an amplifier stage the screen-grid current pulsates similarly to the anode current. If the alternating component of screen-grid current were allowed to pass through R_{g2} (or a divider), the voltage across it would likewise pulsate, and so would the screen-grid voltage. These changes would occur in anti-phase with the alternating voltage at the control grid, and the alternating component of anode current would be reduced. When, on the other hand, the alternating component of screen-grid current is allowed to pass through C_{g2}, the voltage drop produced across R_{g2} is solely due to the direct component of this current, while the voltage drop across C_{g2} which presents a very small impedance is very small. In this way, the screen-grid circuit operates actually at no-load (in terms of alternating current), and $V_{\text{g2 0}}$ remains constant.

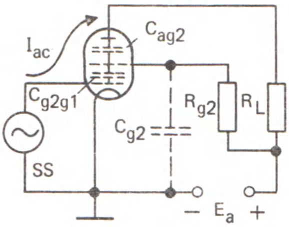


Fig. 15-7

Interelectrode capacitances in a tetrode

When considering the effect of the screen grid, we ought not to confuse the alternating component of screen-grid current I_{g2} with the alternating current passing via the interelectrode capacitances. I_{g2} is part of the electron stream produced by the cathode emission. This current is generated by the triode section of the tube, made up of its cathode, control grid and screen grid. If the tube is at cutoff or the cathode is not hot enough to emit electrons, I_{g2} will be zero. In contrast, the currents passing through the interelectrode capacitances are not electron streams in a vacuum. For example, the capacitive current flowing from the signal source via C_{g2g1} and C_{g2} exists independently of whether the tube is conducting or not or whether the cathode is emitting electrons or not.

15-6 The Characteristics of Tetrodes and Pentodes

Figure 15-8 shows the anode current/grid voltage and screen-grid current/grid voltage characteristics of a tetrode or pentode for two values of anode voltage. Each pair of curves lying close to each other corresponds to a particular value of screen-grid voltage. Even an appreciable change in anode voltage, say, from 200 to 300 volts, will affect the characteristics only slightly. The small change in i_a and i_{g2} caused by a change in anode voltage happens mainly due to the division of space current between the anode and the screen grid. In contrast, a change in screen-grid voltage from, say, 50 volts to 100 volts brings about a marked shift in the characteristics. This shift is roughly proportional to the screen-grid voltage.

The characteristics shown dashed in the figure run lower because the screen-grid current is smaller than the anode current. The curves start from the same origin for both currents – this means that the tube is driven to cutoff at the same time for both its anode current and its screen-grid current. At a lower anode voltage, V_{a1}, the screen-grid curve runs a bit higher owing to the current division. When the ratio v_a/v_{g2} is reduced, the current division ratio k_c, defined as the ratio $i_a/(i_a + i_{g2})$, is also reduced while the cathode current $i_k = i_a + i_{g2}$ remains practically unaffected. Therefore, there is a fall in anode current and about the same rise in screen-grid current.

Tetrodes and pentodes have more nonlinear anode-grid characteristics than do triodes, and the curves plotted for different anode voltages form diverging families. This may be explained as follows. A change in anode voltage affects cathode current only slightly, but it changes the division of space current between the anode and the screen grid. Let, as an example, the change in anode voltage from 200 to 300 volts raise the anode current by 10%. With an anode current of 1 mA, this increase will amount to 0.1 mA, while at 10 mA it will be 1 mA. The screen-grid current will fall by about the same amount. Thus, a rise in anode current leads to a greater divergence among the characteristic curves.

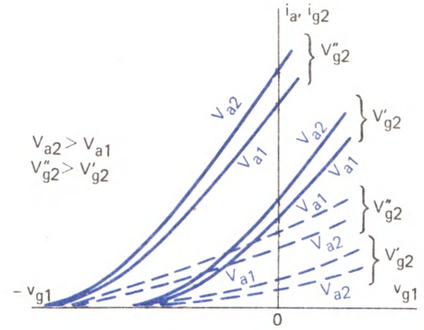


Fig. 15-8

Anode and screen-grid characteristics of a tetrode or pentode

Practical calculations are based on curves for anode, screen-grid and cathode currents plotted at constant voltages at all the grids (Fig. 15-9*a*). Cathode current changes but little with changes in anode voltage, while the characteristics for anode and screen-grid currents may be divided into two regions. In region *I* (the fall-back mode), there is an abrupt rise in anode current and an abrupt fall in screen-grid current in response to small changes in anode voltage because a second potential barrier is formed near the suppressor grid at low values of anode voltage. At $v_a = 0$, no electrons are able to climb over this barrier and fall back to the screen grid. The screen-grid current is a maximum, and the electrons able to reach the anode are only those which have high initial velocities. They constitute a low leakage anode current I_0.

The anode has a strong effect on the second potential barrier. Therefore, even a minute rise in anode voltage leads to a rise in anode current, while the screen-grid current falls because the electrons leaving for the anode do not come back to the screen. As the anode voltage rises, the second potential barrier is reduced in height until all of the electrons passing through the screen grid are able to climb over the barrier, and the intercept mode sets in.

With any further increase in anode voltage, the anode current keeps rising mainly due to the division of current between the anode and the screen. The anode acts on the potential barrier at the cathode through the three grids, and its influence is weakened by a factor of several hundreds or even thousands. Therefore even marked changes in anode voltage can bring about very small changes in anode and screen-grid currents (region *II*). It is these regions of the characteristics that are utilized in the operation of pentodes. High values of the amplification factor and a. c. anode resistance are obtained when the tube is operating within region *II*. This region ought not to be regarded as the temperature-limited (or saturation) region.

Figure 15-9*b* shows a family of anode characteristics for a pentode at $v_{g2} =$ const and $v_{g3} =$ = const. The more negative the grid is, the smaller the anode current is and the lower the characteristics run. As is seen, they are more

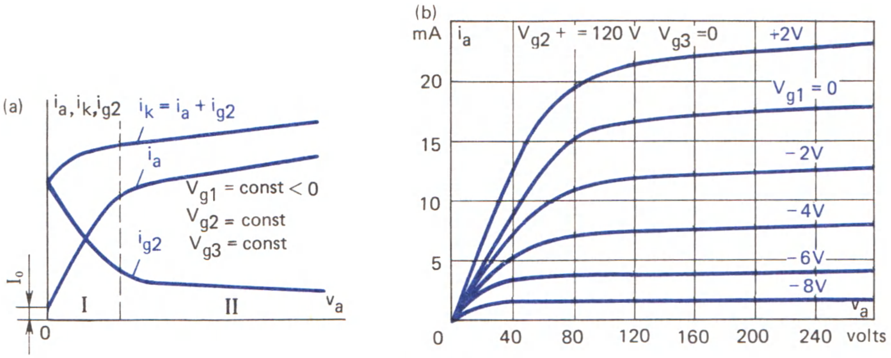


Fig. 15-9

Pentode: (*a*) anode, screen-grid and cathode characteristics and (*b*) a family of anode characteristics

gently sloping and spaced closer together. If we raise the screen-grid voltage, the characteristics will move upwards, and the boundary between regions *I* and *II* (Fig. 15-9*a*) will be shifted to the right. The higher the voltage at the screen grid, the greater the anode voltage that is needed in order for the tube to move from the fall-back to the intercept mode.

When designing pentode-based circuits, a need may arise in the characteristics that show the pentode's currents as functions of the voltage at the suppressor grid often used as a second control grid (Fig. 15-10). This form of control is possible only when the suppressor grid is negative and there is a drastic change in the current division ratio, k_c. As the suppressor grid is made progressively more negative, both k_c and anode current are reduced because the potential barrier at the suppressor grid grows higher and an ever increasing number of electrons fall back to the screen. At a certain negative voltage on the suppressor grid, the tube is driven to cutoff for anode current, while the screen-grid current reaches a maximum value equal to the cathode current. All electrons passing through the screen grid fall back. It may be added that a lower negative voltage is needed to drive the tube to cutoff when the suppressor grid is wound with a finer pitch. Figure 15-10 also shows a plot of cathode current. As is seen, it is only slightly dependent on the voltage at the suppressor grid.

When use is made of only the triode section of a pentode or when a pentode is connected as a triode, the necessary calculations are carried out, using the respective characteristics that do not differ from those of the conventional triode. A pentode is connected as a triode by carrying the screen grid to the anode. If the suppressor grid has a lead, it is likewise connected to the anode.

15-7 The Parameters of Tetrodes and Pentodes

The parameters (or constants) of tetrodes and pentodes are analogous to those of triodes. The transconductance of a tetrode or pentode is given by

$$g_m = \Delta i_a / \Delta v_{g1} \text{ with } v_a, v_{g2} \text{ and } v_{g3} \text{ held constant} \qquad (15\text{-}21)$$

The control grid in tetrodes and pentodes is located in the same way as it is in triodes. Therefore, tetrodes and pentodes have about the same transconductance, that is, units or tens of milliamperes per volt.

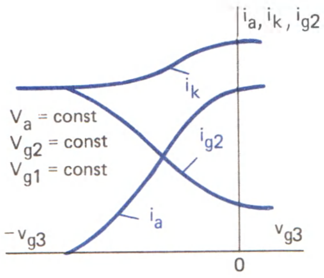


Fig. 15-10

Anode, screen grid and cathode currents as functions of suppressor-grid voltage

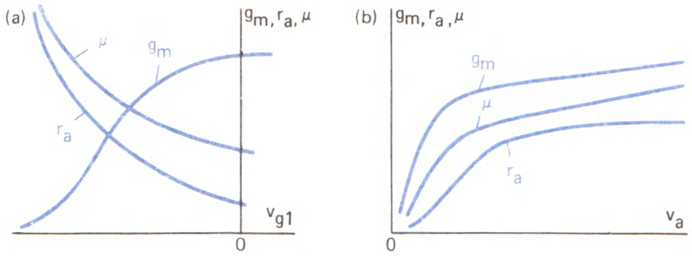


Fig. 15-11

Pentode parameters as functions of (*a*) control-grid voltage and (*b*) anode voltage

The a. c. anode (or dynamic plate) resistance is

$$r_a = \Delta v_a/\Delta i_a \text{ with } v_{g1}, v_{g2} \text{ and } v_{g3} \text{ held constant} \quad (15\text{-}22)$$

Since the action of anode voltage in a tetrode or a pentode is weakened manyfold, r_a may be as high as hundreds or kilohms or even units of megohms and strongly depends on the division of current because when the anode voltage is changed the current changes mainly due to the current division.

The amplification factor is defined as

$$\mu = -\Delta v_a/\Delta v_{g1} \text{ with } i_a, v_{g2} \text{ and } v_{g3} \text{ held constant} \quad (15\text{-}23)$$

and may be several hundreds or even thousands.

The relation $\mu = g_m r_a$ holds as before. The penetration factor D of tetrodes and pentodes is not the reciprocal of their amplification factors because the independent variable is the cathode rather than the anode current:

$$D = -\Delta v_{g1}/\Delta v_a \text{ with } i_k, v_{g2} \text{ and } v_{g3} \text{ held constant} \quad (15\text{-}24)$$

The only parameter that can be found from anode-grid characteristics with sufficient accuracy is the transconductance of a tetrode or pentode.

Changes in the operating conditions bring about appreciable variations in the tube constants because the characteristics of tetrodes and pentodes are strongly nonlinear. As the control grid is made more negative (which is equivalent to a reduction in anode current), the transconductance decreases and the amplification factor increases (Fig. 15-11*a*). Typically, the amplification factor of tetrodes and pentodes also varies with changes in the operating voltages and currents.

Figure 15-12 illustrates how the parameters of a pentode can be derived from its anode characteristics at a specified operating (*Q*-), point. The transconductance is found by two points, *A* and *B*, and the a. c. anode resistance by points *C* and *D*, but the accuracy is low, because the current increment is too small. It is not convenient to determine the amplification factor directly from the characteristics. Knowing g_m and r_a, it can be found by the equation

$$\mu = g_m r_a$$

In the intercept mode, the parameters g_m, r_a and μ are maximal. At low values of anode voltage, that is, in the fall-back mode, the parameters are drastically reduced (Fig. 15-11*b*).

As the control grid is made more negative, the anode characteristics are progressively crowded together, and this corresponds to an increase in r_a and a decrease in g_m (Fig. 15-11*a*).

The parameters of the triode section of a tetrode or pentode, that is, g_{mt}, r_{at} and μ_t can be found by the usual equations, recalling that the screen grid acts as the anode of this triode section. The parameters thus found are similar to those of the usual triode. When a tetrode or a pentode is connected as a triode, its parameters

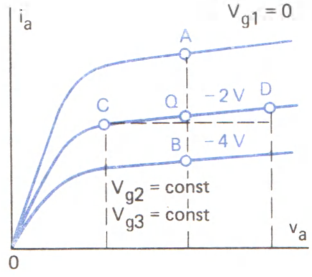


Fig. 15-12

Finding the characteristics of a pentode from anode characteristics differ only slightly from those of the triode section.

When the suppressor grid is used as a control grid, one has to consider the suppressor-grid transconductance and the suppressor-grid amplification factor of the tube:

$g_{m3} = \Delta i_a / \Delta v_{g3}$ with v_a, v_{g1} and v_{g2} held constant (15-25)

$\mu_3 = - \Delta v_a / \Delta v_{g3}$ with i_a, v_{g1} and v_{g2} held constant (15-26)

In calculating the operating conditions and in the practical use of tetrodes and pentodes it is also important to know their absolute maximum ratings, notably the screen-grid dissipation, $P_{g2\,\max}$.

15-8 The Interelectrode Capacitances of Tetrodes and Pentodes

Figure 15-13 shows the circuit of an amplifier stage built around a tetrode. In addition to the interelectrode capacitances existing in a triode, that is, C_{g1k}, C_{ag1} and C_{ak}, there are also the inter-grid capacitance C_{g1g2}, the anode-screen capacitance C_{ag2}, and the screen-cathode capacitance C_{g2k}.

The input capacitance of a tetrode in the dynamic operation (at load) is

$$C_{in\,d} = C_{g1k} + C_{g1g2} + C_{ag1}(1 + k_V) \quad (15\text{-}27)$$

The transfer capacitance C_{ag1} in a tetrode is a fraction of a picofarad. Therefore the term $C_{ag1}(1 + k_V)$ is substantially smaller than the first two terms. Therefore, it is safe to write that

$$C_{in\,d} \approx C_{g1k} + C_{g1g2} \quad (15\text{-}28)$$

As is seen, the capacitance of a tetrode in the dynamic operation (at load) is appreciably smaller than that of a triode.

Example. Compare the input capacitance of a triode stage for which $C_{gk} = 12$ pF, $C_{ag} = 6$ pF and $k_V = 20$, and of a tetrode stage for which $C_{g1k} = 12$ pF, $C_{g1g2} = 10$ pF, $C_{ag1} = 0.02$ pF, and $k_V = 100$.

In the static (no-load) operation, the input capacitance of the triode stage is

$$C_{in} = C_{gk} + C_{ag} = 12 + 6 = 18 \text{ pF}$$

and that of the tetrode stage is

$$C_{in} = C_{g1k} + C_{g1g2} = 12 + 10 = 22 \text{ pF}$$

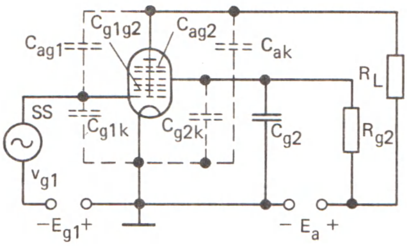


Fig. 15-13
Amplifier stage built around a tetrode

In the dynamic (at-load) operation, the respective figures for the triode stage are

$$C_{in\,d} = C_{gk} + C_{ag}(1 + k_V)$$
$$= 12 + 6(1 + 20) = 138 \text{ pF}$$

and for the tetrode stage

$$C_{in\,d} \approx C_{in} = 22 \text{ pF}$$

The output capacitance of a tetrode is

$$C_{out} = C_{ak} + C_{ag2} \quad (15\text{-}29)$$

which is a bit greater than for a triode (for which $C_{out} = C_{ak}$).

The pentode has 10 interelectrode capacitances. In an amplifier stage, however, the screen and suppressor grids are usually shorted to the cathode for the a.c. component. Therefore, C_{g2k}, C_{g3k} and C_{g2g3} are likewise shorted out. The input capacitance of a pentode in the dynamic operation (at load) is

$$C_{in\,d} \approx C_{in} = C_{g1k} + C_{g1g2} + C_{g1g3} \quad (15\text{-}30)$$

The output capacitance of a pentode is

$$C_{out} = C_{ak} + C_{ag3} + C_{ag2} \quad (15\text{-}31)$$

It follows from the foregoing that the input capacitance of a tetrode or of a pentode is the capacitance between the control grid and all the other electrodes shorted to the cathode for a.c., while the output capacitance is the capacitance from the same electrodes to the anode.

15-9 The Beam Power Tetrode

The *beam power tetrode* or, simply, the *beam tetrode* is a tetrode with characteristics quite similar to those of the pentode. In it, the disadvantage of secondary emission is overcome by setting up a 'potential barrier' for secondary electrons between the screen and the anode.

As compared with the conventional tetrode, the beam tetrode has a greater spacing between its anode and screen, the control and screen grids are wound with the same number of turns and the wires in one are positioned precisely in the shadow of those of the other as seen from the cathode. With this arrangement, the stream of electrons within the tube is concentrated into sheetlike beams (Fig. 15-14). This action is produced by what are called *beam-forming plates* or *rods, BFP_1 and BFP_2*, connected to the cathode. Also, the cathode surface opposite the grid tabs is not given an oxide coating and does not therefore emit electrons.

In the beam tetrode, the electrons follow the paths shown dashed in the figure. Since the beams are kept from spreading sidewise, the density of the space charge is greatly increased. This causes the potential in the anode-screen space to fall. If the anode voltage is lower than that of the screen grid, a potential barrier is formed for the secondary electrons in the screen-anode space of a beam tetrode.

Figure 15-15 shows the distribution of electrons in the beam and the potential distribution in the anode-screen grid space at $v_a < v_{g2}$. Curve *1* holds for both the conventional tetrode and the beam tetrode if it draws a small current and the fall of potential is not enough for a potential barrier to be set up. Curve *2* applies to the beam tetrode with a normal anode current. As is seen, at $v_a = 50$ V and $v_{g2} = 200$ V the potential barrier seen by the secondary electrons is 30 V 'high'. Within the region between $\varphi_{min} = 20$ V and the anode the secondary electrons are acted upon by a retarding field which drives them back to the anode. The secondary electrons are not able to overcome the potential barrier and to reach the screen grid although it is at a higher potential than the anode. In contrast, the primary electrons accelerated to high velocities by the screen-grid voltage are able to climb over this barrier and reach the anode.

In the conventional tetrode, the screen grid breaks up the electron beam and intercepts many electrons. The grid tabs act in a similar manner. Therefore, the electron beam in the conventional tetrode is not dense enough and no potential barrier can be set up for the secondary electrons. A further advantage of the beam tetrode is a reduced screen-grid current which does not exceed 5-7% of the anode current. The electrons travel between the screen-grid wires, and very few of them are intercepted there.

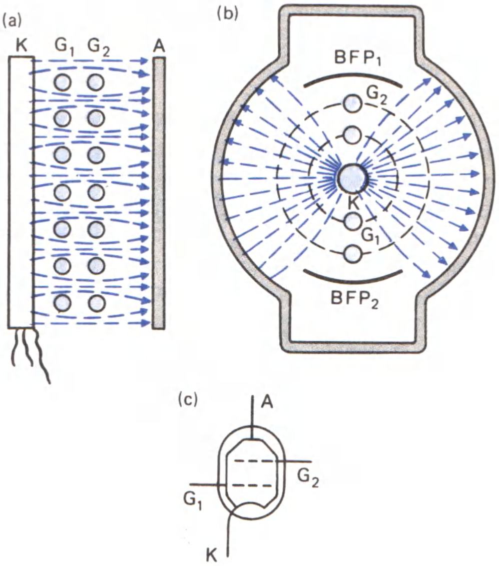


Fig. 15-14

Beam tetrode: structure and graphical (circuit) symbol

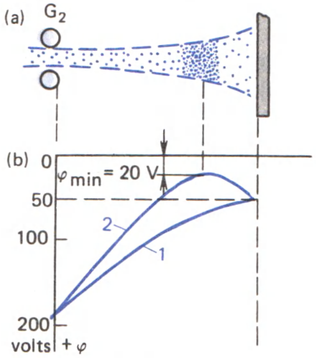


Fig. 15-15

Beam tetrode: (*a*) electron distribution and (*b*) potential distribution

15-10 The Characteristics and Parameters of the Beam Tetrode

The beam tetrode has anode-grid characteristics similar to those of the conventional tetrode or the pentode (see Fig. 15-8). The principal characteristics of the beam tetrode are its anode characteristics (Fig. 15-16). They are similar to those of the pentode but with several differences. For one thing, the transition from region *I* to region *II* is more sharply defined. This is because the anode affects the second potential barrier in the beam tetrode stronger than it does in the pentode (because there are no suppressor-grid wires to interfere with its action). In consequence, region *II* is expanded at the expense of region *I*.

Another difference is that when its control grid is highly negative the beam tetrode displays the effect of secondary emission. Now the cathode current is small, and the space charge is not dense enough to set up a potential barrier that would hold back the secondary electrons. As is seen from the figure, the secondary emission effect gains in strength with decreasing anode current. However, the beam tetrode does not usually operate at low anode voltages and currents. Therefore, the secondary emission effect does not practically manifest itself in the beam tetrode.

The parameters of the beam tetrode can be found by the same equations, Eqs. (15-21) through (15-24), as for the conventional tetrode. The two grids of the beam tetrode have about the same penetration factor, but the control grid is wound with a moderately fine pitch so that the tube has a long grid base, that is, its anode-grid characteristics lie in the region of negative grid voltages. In consequence, the screen grid has likewise a moderately fine pitch and the amplification factor is somewhat smaller than it is for the conventional tetrode. The a.c. anode resistance ranges from tens to hundreds of kilohms. The transconductance is the same as for other tubes, that is, somewhere between units and 10-50 mA V^{-1}. On moving from region *II* into region *I* of the anode characteristics, g_m, r_a and μ of the beam tetrode decrease abruptly.

The beam tetrode has about the same interelectrode capacitances as the conventional tetrode, but C_{ag1} is somewhat greater because the screen grid has a coarser pitch.

In an amplifier stage, the beam tetrode will usually be connected similarly to the conventional tetrode. The screen-grid voltage may be equal to or even a bit higher than the anode voltage (in higher-power stages). In the latter case, the anode voltage ought not to be turned off or the anode circuit ought not to be opened while a full voltage is maintained at the screen grid. Failure to observe this simple rule might lead to an abpurt rise in the screen-grid current, and the screen grid might be overheated.

The beam tetrode is a good substitute for the pentode in power stages. This does not mean to say that pentodes are on their way out of use. Simply, beam tetrodes compare favourably with pentodes in terms of characteristics and their screen grid draws a smaller current. However, they are more difficult to make because their grid must be precisely aligned and beam-forming plates have to be installed. Since beam tetrodes do display the secondary-emission effect at low anode currents, they are not fabricated for low power ratings. The absence of a suppressor grid makes beam tetrodes less versatile than pentodes because the suppressor grid can sometimes be used as a second control grid. Also, by applying a suitable direct voltage to the suppressor grid, it is possible to control the operating conditions of the pentode. Last but not least, pentodes have a higher amplification factor and a lower transfer capacitance.

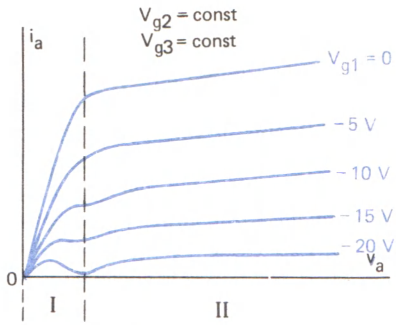


Fig. 15-16

Family of anode characteristics of a beam tetrode

15-11 The Dynamic Operation of Tetrodes and Pentodes

The dynamic operation of tetrodes and pentodes (that is, at load) is usually analysed graphically, using their anode characteristics.

If the useful output power is to be a maximum and the signal distortion is to be kept to a minimum, the load impedance of a tetrode or a pentode must be a fraction of its a.c. anode resistance. This requirement is obvious from reference to Fig. 15-17 which shows the anode characteristics and load lines of a pentode for several values of load resistance (R_{L1}, R_{L2}, and R_{L3}). The operating point on each load line (Q_1, Q_2 and Q) corresponds to a grid cutoff voltage $E_{g1} = -4$ V, and each operating region (A_1B_1, A_2B_2, and AB) corresponds to an alternating grid voltage with an amplitude of 4 V.

When the load resistance is large (R_{L1}), the operating region A_1B_1 is relatively narrow because it partly falls within the region of the fall-back mode. In consequence, the alternating components of anode current and voltage have low amplitudes (peak values). Voltage amplification is insufficient and the useful power is likewise small. With a load resistance equal to R_{L1}, the useful (output) power triangle has a small area. On the other hand, nonlinear distortion is appreciable. To demonstrate, the upper portion of the operating region, Q_1A_1, is substantially smaller than the lower portion of the same region, Q_1B_1. Therefore, the positive and negative half-cycles of alternating anode current and voltage differ markedly in magnitude.

When the load resistance is small (R_{L2}), the operating region (A_2B_2) is extended. The alternating component of anode current now has a large amplitude, but that of the alternating voltage is small because the load line has a very large slope. The useful power is now greater (the power triangle has increased in area), but it is not a maximum, and nonlinear distortion is present again (the upper portion, Q_2A_2, of the operating region is longer than the lower portion, Q_2B_2).

It is possible to choose an optimal load resistance, R_L, such that the operating point will divide the operating region into two equal parts, and nonlinear distortion will therefore be a minimum. With this value of load resistance, R_L, the upper portion, QA, of the operating region is equal to the lower portion, QB. The alternating component of anode current now has positive and negative half-cycles of the same amplitude (peak value), and V_{mR} is substantially higher than it was in the two previous cases. There is an increase in the useful power as well – the power triangle has an increased area. The optimal load line runs steeper than the static load line. This implies that R_L is substantially smaller than r_a. For most pentodes and beam tetrodes the optimum load resistance is

$$R_L = 0.05r_a \text{ to } 0.2r_a \tag{15-32}$$

In rough terms, it is assumed that R_L should be equal to about $0.1r_a$. Any departure from the optimal value of R_L causes a decrease in the useful power (albeit, insignificant), and nonlinear distortion is exacerbated. The plots of Fig. 15-18 show how P_{out} and the harmonic ratio k_h depend on the load ratio defined as R_L/r_a.

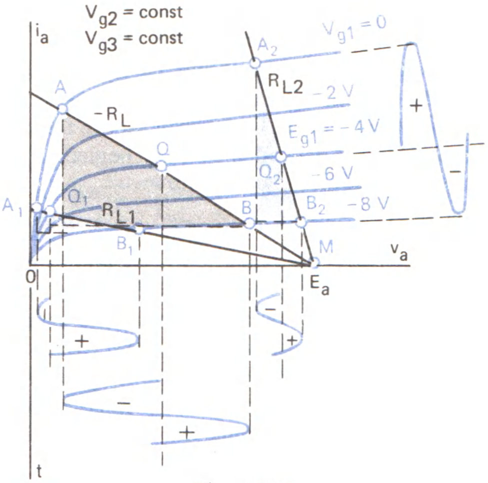


Fig. 15-17

Graphical analysis of pentode operation as an amplifier for several values of load resistance

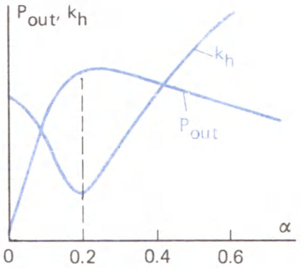


Fig. 15-18

Useful power and harmonic content as functions of the load factor

The optimal location for the operating (Q-) point is found by rotating a rule about the point M where $v_a = E_a$ (see Fig. 15-17). The rule should be positioned so that the characteristics corresponding to the chosen bias and the ends of the operating region intercept equal segments, QA and QB, on the rule. Then the value of R_L is found by dividing E_a by the current corresponding to the intersection of the load line and the axis of ordinates.

If the load resistance, R_L, is large only for the a.c. component and is very small for direct current (this happens in a transformer-coupled or tuned amplifier), the load lines for different values of R_L will intersect at the operating (Q-) point and not point M. In order to determine the optimal operating conditions in such a case, the rule should be rotated about point Q until the operating region is divided into two equal halves.

The stage gain in the case of tetrodes and pentodes is found in the same manner as for triodes (see Sec. 14-4). If we recall that in the case of tetrodes and pentodes we may neglect R_L in comparison with r_a, we obtain

$$k_V \approx g_m R_L \tag{15-33}$$

In other words, the stage gain is roughly proportional to the transconductance of the tube. The greater the transconductance, the greater the amplification. Equation (15-33) yields an overrated value of the stage gain k_V; still, it is widely used for practical calculations. It is to remember that it holds only when R_L is a small fraction of r_a. For triodes whose R_L is of the same order ot magnitude as r_a, this equation may not be used. In Eq. (15-33) it is convenient to express g_m in mA V^{-1} and R_L in kΩ. For example, if $g_m = 2$ mA V^{-1} and $R_L = 100$ kΩ, then $k_V \approx 2 \times 100 = 200$.

Since for tetrodes and pentodes $R_L \ll r_a$, the Norton equivalent for the anode circuit of an amplifier stage may omit r_a because the a.c. anode resistance and the load resistance are connected in parallel. This yields a simplified equivalent circuit such as shown in Fig. 15-19.

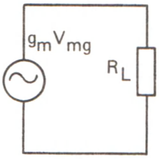


Fig. 15-19

Simplified a.c. equivalent anode circuit of a pentode (tetrode)

15-12 Variable-Mu Tubes

The large amount of amplification produced by pentodes in the r.f. amplifiers of receivers is an asset so far as weak signals are concerned. With a strong incoming signal, however, a prohibitively heavy nonlinear distortion might arise. What is needed in such situations is a tube whose amplification factor could conveniently be varied in inverse proportion to the signal strength. When thus built, the tube has what is known as a *variable-mu characteristic*. As is seen from the plot of Fig. 15-20*a*, the characteristic has a long grid base, or a *remote-cutoff point*. Hence, another name for such tubes is *remote-cutoff tubes*.

The desired characteristic is obtained by varying the pitch of the control grid winding: a tube which has a closely wound grid has a fairly high transconductance, whereas when the grid is 'openly' wound its g_m is low. In a variable-mu (or remote-cutoff) tube the small part in the middle of the control grid has a coarse pitch, and the remainder has a fine pitch, as shown in Fig. 15-20*b*. With a large negative bias the tube is cut off within the closely wound portion of the grid and only the openly wound part of the grid determines the value of anode current (curve *1*). This characteristic corresponds to a low transconductance and a high cutoff voltage. The stage gain $k_V \approx g_m R_L$ is low. With a small negative bias the closely wound section of the grid is the effective part which determines the value of anode current. This grid section is associated with curve *2* which corresponds to an appreciable transconductance but a low cutoff voltage. The high transconductance results in a high stage gain. The overall characteristic of the tube (the full line) is obtained by adding the currents defined by curves *1* and *2*.

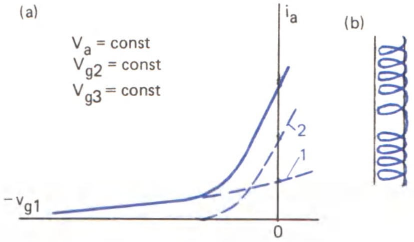


Fig. 15-20

Variable-mu tube: (*a*) characteristics and (*b*) control grid

For weak signals, the operating point is chosen to lie within the steeply sloping part of the curve (point Q_1 in Fig. 15-21), whereas for strong signals a more negative bias is chosen and the operating point is located within the gently sloping part of the curve (point Q_2). The swing of anode current is about the same in each case. Thus, signals varying in amplitude can be received at about the same volume while avoiding the distortion that might have been caused by an excessively strong signal. The operating point is positioned as required automatically. After detection (rectification), stronger signals produce a direct voltage which is applied as an additional grid bias to a variable-mu tube and shifts the operating point within the more gently sloping part of the characteristic (that is, one with a low transconductance). This arrangement is known as *automatic gain control*, or AGC for short.

15-13 Basic Types of Tetrodes and Pentodes

Several designs of tetrodes are commercially available for use in receivers and amplifiers. Some tetrodes are specifically intended for use as high-power modulator tubes in pulse working and high-power transmitting tubes. Beam tetrodes are used in the final stages of a.f. amplifiers and also in oscillators and transmitters. Low-power beam tetrodes are not manufactured because at low anode currents they suffer from the consequences of secondary electron emission (see Sec. 15-8).

Pentodes are the most commonly used tubes. Receiving-amplifying pentodes are classed into low-power for use at high and low frequencies, and medium-power for use at low frequencies. Medium- and high-power pentodes are also used in oscillators and transmitters. In fact, special-purpose transmitting pentodes constitute a large group in its own right.

The earlier designs of pentodes had the control grid connection at the top of the envelope and the anode connection at the base. State-of-the-art pentodes are more convenient because all of the electrodes have their connections brought out to the base, with the anode and the control grid usually connected to the diametrically opposite pins. The electrode structure includes shields to minimize the effect of the anode-to-control grid capacitance. There is a metal shield inside the envelope and in the alignment pin, which is connected to chassis ground (Fig. 15-22). In bantam tubes which have no alignment pin, there is a shield in the central opening of the tube socket. Such shields greatly reduce the transfer capacitance.

Wide use is made of low-power pentodes, notably maniaturized and bantam designs.

Pentodes for the final stages of a.f. amplifiers have all of their electrodes brought out to the base unshielded because the low transfer capacitance existing at audio frequencies does not affect the tube performance.

All transmitting pentodes usually have a suppressor-grid connection because it is sometimes practised to apply a positive voltage to this grid in telegraph transmitters as a way of enhancing the useful power output, while in telephone transmitters this grid is used for modulation.

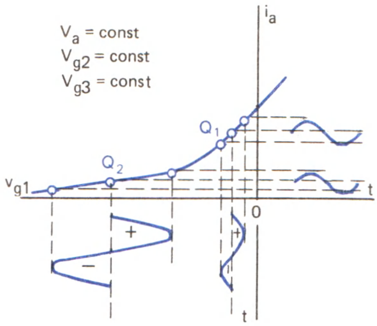


Fig. 15-21
Variable-mu tube and its amplification effect

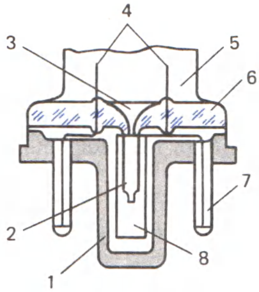


Fig. 15-22
Electrode leads and connections in a single-base tube: (*1*) key (alignment pin); (*2*) exhaust tube; (*3*) shield; (*4*) electrode leads; (*5*) envelope; (*6*) glass; (*7*) pin; (*8*) shield

15-14 Miscellaneous Tubes

The superheterodyne receiver always includes at least one stage for changing the frequency of the incoming signal to the fixed frequency of the main intermediate-frequency (IF) amplifier in the receiver. This *frequency-changing process* is accomplished by *beating together* (*heterodyning*) a locally generated oscillation and the incoming signal frequency. The local oscillation may be generated by some elements within the *frequency-changing tube* – in such cases, the frequency-changing tube is commonly called a *converter tube*, usually fitted with two control grids. A common multigrid converter tube is the *heptode* (or seven-electrode tube). It has five grids, and so it is more commonly called a *pentagrid tube*. The triode section of a pentagrid converter tube, made up of the cathode and the first two grids, operates as the *local oscillator* supplying the auxiliary frequency required for frequency conversion. The second grid operates as the anode of the triode and as a screen grid. The third grid does the job of a second control grid and is usually called the *signal grid* because the incoming signal is applied to it. The fourth and fifth grids are the usual screen and suppressor grids as they are in a pentode.

In some cases, the local-oscillator signal is supplied by a separate tube, and the heptode is only used to mix the incoming signal and the local-oscillator output frequency. Quite aptly, such a tube is called a *mixer tube*. Unfortunately, heptodes show a poor performance at wavelengths shorter than 20 m, and so they may be encountered only in the older makes of equipment.

At one time, the job of frequency changing was also done by *hexodes* (or six-electrode tubes) which differ from heptodes in that they have no suppressor grid, and by *octodes* (or eight-electrode tubes) in which the second grid operated as the plate of a triode and the third grid acted as a screen separating the local-oscillator and signal sections of the tube.

Use has also been made of *multi-unit tubes* where a single envelope contains several electrode structures. Such tubes reduce the size of the equipment and simplify the wiring problems. In diagrams, it is usual to show the heater and cathode of only one electrode structure so as to make the graphic presentation less crowded. Many of such tubes, especially those intended for service at high frequencies, have shields to avoid parasitic capacitive coupling between the various electrode structures.

Receivers, electronic instruments and tape recorders often use 'magic eyes' – electron-ray tubes which indicate visibly on a fluorescent target the effects of changes in a control-grid voltage applied to the tube. In receivers, they are used as *tuning indicators*; with them, any desired station can be tuned in with the receiver's volume control set to zero ('silent tuning'). In instruments and tape recorders, they indicate the voltage level used. A typical electron-ray tube, or a magic eye, consists of an amplifying triode and a triode indicator system in which the anode is used as a fluorescent target. The magic eye operates so that its dark segment becomes wider or narrower according as it is bombarded by a stronger or a weaker electron stream.

The transconductance of amplifying tubes can be enhanced in ways other than decreasing the grid-to-cathode spacing (see Chap. 13). For example, what is known as the *space-charge tetrode* uses an additional grid placed between the control grid and the cathode and held at some positive potential. This additional grid serves to produce a potential barrier (a space charge) close to the control grid. As a result, the control grid produces a stronger action on the barrier and, in consequence, on the electron beam. The transconductance is as high as 25 mA V^{-1}, but there is an appreciable waste of power due to the current drawn by the additional grid.

Secondary-emission tubes have one or several dynodes, that is, electrodes having the primary function of supplying secondary-electron emission and held at a less positive potential than the anode. The primary electrons emitted by the cathode would strike the dynode and knock out many more secondary electrons. This principle can boost the transconductance to several hundred milliamperes per volt.

Some time ago, V. N. Avdeyev of the Soviet Union proposed to use rod electrodes in tubes. With them, the tube needed a lower filament power and a smaller anode supply source, and had smaller interelectrode capacitances and screen-grid current. Also, they were mechanically robust and reliable. Unfortunately, they could offer a relatively low transconductance.

At one time there was a good deal of interest in *nuvistors*, extremely small receiving-amplifying metal-ceramic triodes and tetrodes. They are highly reliable and economical, can be manufactured on an automatic production line, and have a highly consistent performance from one unit to another. Nuvistors are insensitive to shocks, jarring and vibration, can operate at temperatures as high as 200°C. Some nuvistors come with cylindrical electrode terminals which can readily be mated with coaxial resonant circuits, and can operate at frequencies up to 2 GHz.

Chapter Sixteen

Cathode-Ray Tubes

16-1 General

Cathode-ray tubes (CRTs) are widely used as indicators (or displays) in radar, in oscilloscopes, as picture and camera tubes in television, as storage tubes, electron-beam switches, in electron microscopes, as electron image converters, and elsewhere. In all of these applications, a narrow beam of electrons is produced and controlled by an electric field or a magnetic field or by both. Most CRTs serve to produce visible images or patterns on a fluorescent screen, thereby permitting the visual observation of electrical signals. This chapter will be concerned with the most commonly used CRO tubes and TV picture tubes closely associated with the CRTs used by radar and sonar equipments.

The main electrical difference between CRT types lies in the means employed for focusing and deflecting the electron beam. The beam may be focused and/or deflected either electrostatically or magnetically because a stream of electrons can be acted upon either by an electrostatic or a magnetic field. Depending on the material used for the screen, a CRT may present a green, orange or yellow-orange colour (for visual observation), a blue colour (to take photographs of the patterns displayed), and white-and-black or tri-colour (for TV). A further distinction is in the time during which the screen stays fluorescent after electrons cease to bombard it – this is known as the *screen persistence*. CRTs may further be classed according to the material used for their envelopes (glass or metal-glass), size, and some other features.

16-2 The Electrostatic CRT

Electrostatic CRTs, that is, ones in which the electron beam is focused and deflected by an electrostatic field, are especially widely used in oscilloscopes.

Figure 16-1 shows the basic arrangement of a simple electrostatic CRT along with its graphical symbol. The envelope is a combination of a cylinder and a cone or of two cylinders differing in diameter. The inside surface of the CRT faceplate is given a coat of a *phosphor*, that is, a material which fluoresces when bombarded by electrons. The envelope encloses an electrode structure with leads taken from the individual electrodes usually to the various pins at the base. In the figure, the leads are shown directly passing through the envelope for simplicity.

The cathode (K) is usually an oxide-coated indirectly heated type in the form of a cylinder that encloses a filament or heater. The oxide is applied to the bottom of the cylinder. The cathode is surrounded by a cylindrical control grid (G), or modulator, which has a small hole in its front for the passage of the electron stream. The function of this electrode is to control the density of the electron stream and to pre-focus it into a narrower beam. The grid is usually held at a negative voltage of several tens of volts. As the voltage applied to the grid is raised, an ever greater number of electrons is caused to travel back to the cathode. At some negative voltage at the grid the CRT is cut off, or *blanked*.

Next in order are the anodes. In the simplest cases, there are two of them. The *second accelerating anode* (A_2) is held at a voltage of 500 V to several kilovolts (sometimes as high as 10-20 kV), and the *first accelerating anode* (A_1) is maintained at a voltage which is a small fraction of that at the second. The anodes enclose perforated partitions (called diaphragms). The accelerating anodes bring the electrons up to

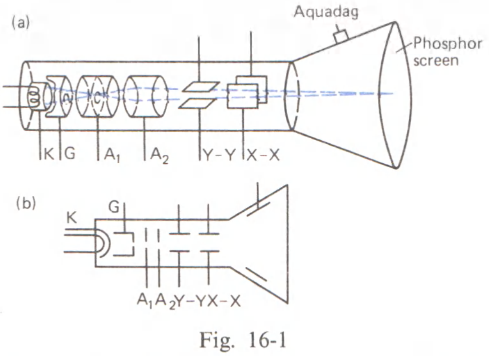


Fig. 16-1

Electrostatic cathode-ray tube: structure and graphical (circuit) symbol a considerable velocity. The electron beam is finally focused by a uniform electrostatic field between the anodes (where an additional focusing electrode may be placed) and by the diaphragms. More elaborate electrode structures may have a greater number of electrodes.

The electrodes which have been described up to this point (the cathode, the control grid, and the anodes) constitute what is known as an *electron gun* which produces free electrons and focuses them into a slender, concentrated, rapidly travelling stream for projecting onto the viewing screen.

To make the CRT useful, means must be provided to deflect the electron beam along two axes at right angles to each other. With electrostatic deflection, this is done by electrostatic deflection plates, one pair (*X-plates*) to deflect the beam horizontally and the other pair (*Y-plates*) to deflect the beam vertically. The voltage applied to the plates produces an electrostatic field which deflects the beam towards the positively charged plate. The field set up by the plates is a transverse one for the electrons. In such a field, the electrons travel along parabolic paths; on leaving the field, they keep moving on in straight lines. In this way, the beam receives a net angular deflection. The higher the voltage applied to the plate, the stronger the deflection. As a result, the luminous spot produced on the viewing screen by the impinging beam moves farther away from its centre.

One plate in each pair is sometimes connected to the chassis ground, and so it is at zero potential. The plates are then said to be *unbalanced*. To avoid an electrostatic field affecting the electrons between the second anode and the chassis ground, it is usual to ground the second anode as well. Then, with no voltage applied to the deflection plates, no field will exist between them and the second anode, and the electron beam will remain unaffected.

Since the second anode is grounded, the cathode held at a high negative potential equal to that at the second anode must be well insulated from the chassis ground. It is dangerous to touch the cathode, grid and filament leads when the CRT is turned on.

Because extraneous electric and magnetic fields might affect the electron beam, the CRT is often enclosed in a mild-steel shield.

The luminescence of the fluorescent screen is caused by the excitation of the phosphor as electrons strike it and is called the *cathodoluminescence*. As the electrons impinge on the screen, they give up their energy to the luminescent *phosphor* atoms so that one of their electrons moves to an orbit farther away from the nucleus. On falling back to its former orbit, the excited electron emits a quantum of radiant energy (a photon), and this is seen as a glow.

Electrons hitting the screen may charge it negatively, thereby setting up a field that might retard the incoming electrons. This might cut down the screen brightness or even prevent electrons from reaching the screen altogether. Therefore, there is a need to conduct the negative charge away from the screen. For this purpose, an *Aquadag* coating (a conductive graphite material) is applied to the inside of the envelope and connected to the second anode. The secondary electrons knocked out of the screen by primary electrons are attracted by the Aquadag coating so that the screen is usually at a potential very close to that of the Aquadag. In some CRTs, the Aquadag coating has a terminal (*AQ* in Fig. 16-1) so that it can be used as a *postdeflection* or a *third accelerating anode* operated at a higher voltage. The term 'postdeflection' refers to the fact that the electrons are additionally accelerated after they have been acted upon by the deflection plates.

The Aquadag coating also serves as an electrostatic shield as it prevents the formation of a negative charge on the envelope walls by the impinging electrons. This electrostatic charge might produce stray fields that would upset the normal operation of the CRT. Without any Aquadag coating, secondary electrons would have moved from the screen to the deflection plates and the second anode.

The electrodes of a CRT are usually set up on metal supports and insulators carried by a glass stem.

16-3 The Supply Circuits of the CRT

The supply circuits of a typical CRT are shown in Fig. 16-2. Direct voltages for the electrodes are fed by two rectifiers, E_1 and E_2. The first rectifier must supply a very high voltage (hundreds or even thousands of volts) at a current of not more than several milliamperes. The E_2 source is designed for a voltage which is a fraction of that supplied by E_1. This source also feeds the other stages operating together with the CRT. Therefore, it supplies a current of several tens of milliamperes.

The electron gun draws its power via a voltage divider made up of resistors R_1, R_2, R_3, and R_4. Their resistances are chosen to be hundreds of kilohms so that the divider could only draw a small current. The CRT itself draws a small current too, mostly tens or hundreds of microamperes.

The potentiometer R_1 is used as the BRIGHTNESS control. It controls the negative voltage applied to the control grid from the right part of R_1. An increase in the absolute value of this voltage reduces the number of electrons in the beam, and so the brightness of the screen is reduced, too.

The potentiometer R_3 is the FOCUS control. It controls the voltage applied to the first anode so that the difference of potential between the two anodes is changed, and so is the field strength between them. If, say, we bring down the potential at the first anode, the difference of potential between the two anodes will rise, the field will gain in strength, and its focusing action will be enhanced. Because the first-anode voltage V_{a1} need not be reduced to zero or increased to that at the second anode, V_{a2}, the divider includes resistors R_2 and R_4.

The second-anode voltage V_{a2} is only slightly lower than E_1 (by the voltage drop across R_1). It is to be remembered that the velocity of the electrons leaving the gun depends solely on the second-anode voltage and is independent of the voltages at the control grid and the first anode. Some of the electrons are intercepted by the anodes, especially if the latter are fitted with diaphragms. Therefore, the anode circuits draw currents of a fraction of a milliampere, which have their path completed through the E_1 supply source. For example, the electrons constituting the current drawn by the first anode move from cathode to anode, then via the right-hand portion of R_3 and R_4 to the " + " terminal of the E_1 source whence they move through the source and R_1 to the cathode.

The potentiometers R_5 and R_6 connected to the E_2 source serve to position the luminous spot on the screen initially. Their wiper arms are connected through high-value resistors R_7 and R_8 to the deflection plates. Resistors R_9 and R_{10} of the same value are used to set the zero-potential point returned to the chassis ground. The potentiometers R_5 and R_6 have a potential of $+ 0.5E_2$ at one end and a potential of $- 0.5E_2$ at the other, while their centre points are at zero potential. When the wiper arms of R_5 and R_6 are in their mid-positions, the voltage across the deflection plates is zero. Moving the wiper arms away from their centre positions will apply any desired voltage to the deflection plates, thereby causing the luminous spot to deflect vertically or horizontally to any point on the screen.

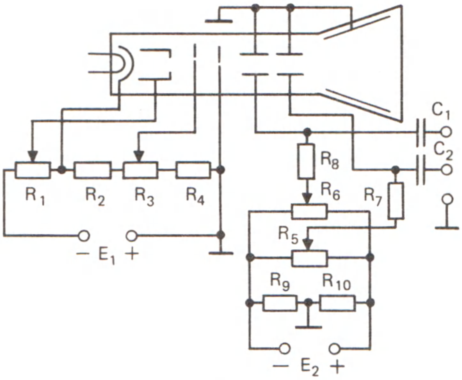


Fig. 16-2

CRT operating from two supply sources

The deflection plates also receive an alternating voltage (say, the voltage waveform of interest when a CRT is used in an oscilloscope) via d. c. blocking capacitors C_1 and C_2. Without these capacitors the deflection plates would be shunted for d. c. by the internal resistance of the a. c. (signal) source. Should this internal resistance be low, the direct voltage across the deflection plates would be drastically reduced. Also, the unknown signal may sometimes contain a direct voltage that ought not to be applied to the deflection plates. In many cases, it is likewise undesirable for the direct voltage used in the deflection-plate circuits to enter the a. c. (signal) source.

The function of the resistors R_7 and R_8 is to raise the input resistance of the deflection system seen by the a. c. (signal) source. Without them, the signal source would be loaded into the substantially smaller resistance supplied by R_5, R_6, R_9, and R_{10}. On the other hand, R_7 and R_8 do not bring down the direct voltage applied to the deflection plates because no direct currents are allowed to flow through them.

The useful current is the electron-beam current. The electrons that constitute it travel from the cathode to the fluorescent screen and knock out of it secondary electrons which move towards the Aquadag coating and farther on towards the " + " terminal of the E_1 source, through its internal resistance and the potentiometer R_1 to the cathode.

Other schemes may be used to energize a CRT, say, from a single H. V. supply.

16-4 The Electron Gun of an Electrostatic CRT

The electron gun of a CRT is essentially an *electron-optical system* made up of a series of electrostatic *electron lenses.* Each lens is formed by a nonuniform electrostatic field which bends the paths of electrons in much the same way as an optical lens causes the refraction of light rays. In addition to bending the travel paths of electrons, electrostatic lenses can accelerate or decelerate the electrons.

The simplest electron gun consists of two lenses. The first or *prefocusing lens* is made up of the cathode, control grid and first anode. Figure 16-3 shows the field in this part of the electron gun. The equipotential surfaces are shown as full lines, and the lines of force as dashed lines. As is seen, some of the lines of force originating at the first anode run towards the space charge near the cathode, and the remainder terminates at the control grid which is more negative than the cathode. Arbitrarily, the line BB' divides the field into two parts. The left-hand part focuses the electron stream and accelerates the electrons. The right-hand part of the field accelerates the electrons still more but somewhat disperses them (causes the beam to spread sideways). However, the spreading action is weaker than the focusing action because the electrons move at a higher velocity in the right-hand part of the field.

This field is not unlike a system of a converging (or positive) lens and a diverging (or negative) lens. The converging lens (also known as the collecting lens) is stronger than the diverging lens, so on the whole the system produces a focusing action. However, the electron streams travel by laws different from those that govern the refraction of light beams in optical lenses.

Figure 16-4 shows the trajectories for the outer electron beams leaving the cathode. The electrons move in curved paths. Their streams are focused and intersect in a small region called the *first cross-over* and located most often between the control grid and the first anode.

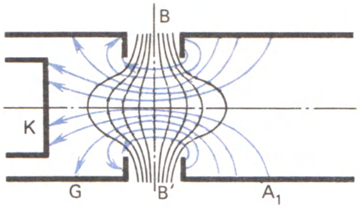


Fig. 16-3
First lens of an electron gun

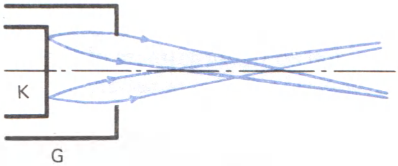


Fig. 16-4
Travel paths of electrons in the first lens of an electron gun

The first lens has a short focal length because the electrons travel through it at relatively low velocities and their paths are curved appreciably.

When the control grid is made progressively more negative, the potential barrier near the cathode builds up, and an ever decreasing number of electrons are now able to climb over it. There is a reduction in the cathode current and, as a consequence, in the electron-beam current and in the brightness of the luminous spot on the screen. The potential barrier grows in height to a lesser extent in the central part of the cathode because this portion of the barrier is more strongly affected by the accelerating field that finds its way from the first anode through the central hole in the control grid. At some negative voltage at the control grid the potential barrier at the edges of the cathode rises so high that no electrons can climb over it. Only the central part of the cathode remains effective. Any further increase in the negative voltage reduces the effective area of the cathode until it becomes equal to zero – the CRT is said then to be blanked. Thus, brightness control is associated with changes in the effective area of the cathode.

Now let us see how the electron beam is

focused by the second lens, that is, the system of the two anodes (Fig. 16-5*a*). Line *BB'* divides the field between the anodes into two parts. On entering the left-hand part, the divergent electron stream is focused while in the right-hand part of the field it is caused to spread or diverge. The spreading action is weaker than the focusing action because the electrons moving in the right-hand part have a higher velocity than they do in the left-hand part. The entire field is not unlike an optical system made up of a converging (positive) and a diverging (negative) lens (Fig. 16-5*b*). Because the electrons move in the field between the two anodes at high velocities, the system has a long focal length. This is what is required because the electron beam has to be brought to a sharp focus at the screen positioned a fairly large distance away.

When the potential difference between the anodes is increased (by reducing, say, the voltage at the first anode), the field gains in strength and produces a stronger focusing action. Focusing could in principle be controlled by varying the voltage at the second anode but this is inconvenient because this would change the velocity of the electrons leaving the gun, thereby causing the light spot on the screen to vary in brightness, and would affect the deflection of the beam by the plates.

A drawback of the electron gun we have just discussed is the interaction between the brightness and focusing controls. Changes in the potential at the first anode affect the brightness because the field of the first anode affects the potential barrier near the cathode. Changes in the potential at the control grid shift the first cross-over along the tube's axis, thereby impairing the sharpness of the focus. Also, manipulation of the brightness control changes the first-anode current. Since, however, the first-anode circuit contains high-value resistors, its voltage is likewise changed, and this leads to defocusing. Changes in the second-anode current do not affect focusing because the second-anode circuit contains no resistors, so its voltage cannot change.

Better electron guns have one more electrode between the control grid and the first anode, called the *screen grid* (or, by some authors, the *preaccelerating anode*) (Fig. 16-6). It is connected to the second anode and held at a constant potential. Owing to this screen grid, changes in the potential at the first anode in response to focus control do not practically affect the field at the cathode.

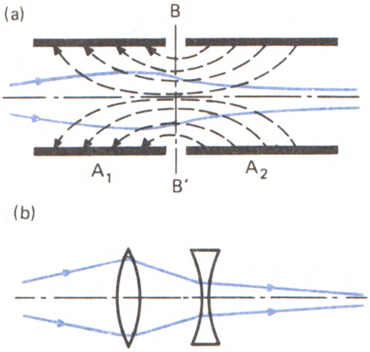


Fig. 16-5

Second focusing lens of an electron gun and its optical analogy

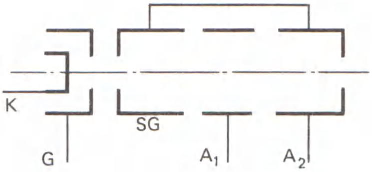


Fig. 16-6

Electron gun and the screen and accelerating electrodes

The focusing system made up of the control grid, the screen grid and the first anode operates as follows. The field between the first and second anodes is as shown in Fig. 16-5*a*. It focuses the electron stream as already explained. The field between the screen grid and the first anode is a nonuniform one, similar to that between the anodes (see Fig. 16-5*a*), but it is retarding rather than accelerating. The electrons entering this field as a divergent stream are spread sidewise in the left-hand part of the field and focused in the right-hand part. The focusing action is stronger because the electrons moving in the right-hand part have a lower velocity. In this way, the stream is focused between the screen grid and the first anode. The lower the potential at the first anode, the stronger the field and the better the focusing action.

To minimize the effect of the BRIGHTNESS control on FOCUS, the first anode has no diaphragm (Fig. 16-6). It does not intercept electrons, so it draws no current.

State-of-the-art electron guns produce a light spot which is 0.001 or 0.002 of the screen's diameter.

16-5 Electrostatic Beam Deflection

The electron beam in a CRT and the luminous spot on its viewing screen are deflected in proportion to the voltage applied to the deflection plates. A measure of the physical displacement of the beam (that is, as measured on the screen) is given by what is called the *deflection sensitivity* of a CRT. If we denote the vertical displacement of the beam (or, which is the same, the luminous spot) as y, and the voltage applied to the Y-plates as V_Y, we may write

$$y = S_Y V_Y \qquad (16\text{-}1)$$

where S_Y is the vertical deflection sensitivity of the CRT.

Similarly, for the horizontal deflection of the beam

$$x = S_X V_X \qquad (16\text{-}2)$$

Thus, the deflection sensitivity of a CRT is the ratio between the physical displacement of the beam as measured on the screen and the applied deflection voltage:

$$S_X = x/V_X$$

and (16-3)

$$S_Y = y/V_Y$$

In other words, the deflection sensitivity of a CRT is the deflection of the beam per volt of deflection voltage. It is usually expressed in millimetres per volt. Many CRO manufacturers quote a quantity which is the reciprocal of the deflection sensitivity and is called the *deflection factor*, expressed in volts per millimetre. Sometimes, the deflection factor is referred to as the deflection sensitivity as well.

Equations (16-3) do not mean to say that the deflection sensitivity is inversely proportional to the deflection voltage. If we increase V_Y several times, the vertical beam displacement y will increase by the same factor, but the value of S_Y will remain unchanged. Thus, S_Y is independent of V_Y. The deflection sensitivity of typical CRTs ranges between 0.1 and 1 mm V^{-1}. It is a function of the operating conditions and tube geometry (Fig. 16-7):

$$S = l_{dp}l/2dV_{a2} \qquad (16\text{-}4)$$

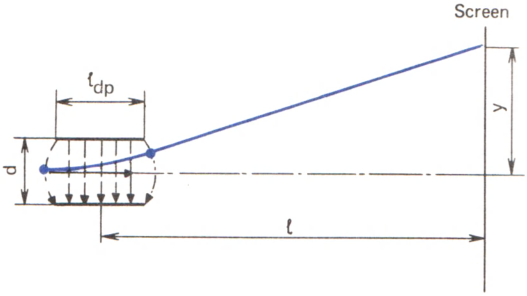


Fig. 16-7
Electrostatic beam deflection

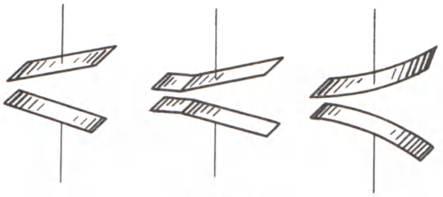


Fig. 16-8
Deflection plates

where l_{dp} = length of the deflection plates
l = spacing between the plates' centre and the screen
d = spacing between the plates
V_{a2} = voltage at the second anode

Equation (16-4) is easy to explain. With an increase in l_{dp}, each electron has to travel a longer distance in the deflecting field and its displacement increases. Given the same angular displacement, the deflection of the light spot on the screen increases with increasing l. If we increase d, the field between the plates will be weaker, and this will lead to a smaller displacement of the light spot. An increase in V_{a2} leads to a reduced deflection because the electrons travel through the field between the plates at a higher velocity.

Let us see how we could increase the sensitivity on the basis of Eq. (16-4). An increase in l is undesirable because too long a tube is inconvenient in use. If we increase l_{dp} or decrease d, we will fail to secure a large beam deflection because the beam will hit the plates. To avoid this, the plates are bent and placed relative to each other as shown in Fig. 16-8. Unfortunately, this entails a reduction in the brightness of the spot, which is undesirable in many cases, especially when the beam is deflected over the screen at a high rate. A reduction in the anode voltage degrades the sharpness of focus as well. At a higher value of V_{a2}, the electrons inside the tube travel at high velocities, and the mutual repulsion of electrons manifests itself to a lesser degree, and their paths in the electron gun make smaller angles with the tube's axis. These paths are called *paraxial trajectories*. They improve the focusing of the beam and reduce distortion in the image displayed on the screen.

The reduction in brightness (beam intensity) secondary to a fall in V_{a2} is made up for in CRTs using *post-deflection acceleration*. In these tubes, the electron gun accelerates electrons to an energy of not more than 1-1.5 keV. Thus accelerated, the electrons travel past the deflection plates and enter the accelerating field set up by a third anode which is the conductive Aquadag coating applied to the inside of the envelope and insulated from the underlying material connected to the second anode (Fig. 16-9*a*). Owing to this arrangement, $V_{a3} > V_{a2}$. The field between these two anodes forms a lens which accelerates electrons. At the same time, it bends the electron paths somewhat so that the sensitivity is reduced a little and there is a slight distortion of the image. These shortcomings are avoided to a considerable degree by multiple post-deflection acceleration provided by a series of conductive rings at which the voltage increases in progression such that $V_{a4} > V_{a3} > V_{a2} > > V_{a1}$ (Fig. 16-9*b*).

The stray capacitance that might exist between the X- and Y-plates is minimized sometimes by passing their leads directly through the glass envelope and by placing a shield between the plates. This purpose may alternatively or jointly be served also when the X-plates are located away from the Y-plates. Because of the different distance from the plates to the screen in such an arrangement, the sensitivity along the X-axis may somewhat differ from that along the Y-axis.

Distortion may further be produced if the deflection voltage is varying at a very high frequency, because the transit time of electrons in the field set up by the deflection plates becomes comparable with the period of oscillation of the deflection voltage. During this time interval the voltage at the plates changes markedly and may even take an opposite sign. This form of distortion can be minimized by using short plates and higher accelerating voltages.

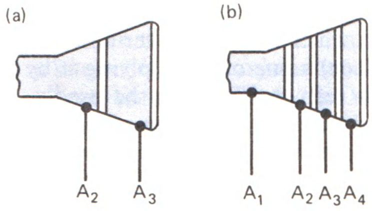


Fig. 16-9
Post-deflection accelerating anodes

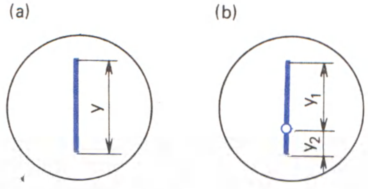


Fig. 16-10
Measuring an alternating voltage with a CRT

A further effect of increasing frequency is the greater effect of the self-capacitance of the deflection plates.

State-of-the-art oscilloscopes for use at microwave frequencies employ far more sophisticated deflection systems.

16-6 Measurement and Visual Observation of Alternating Voltages with a Cathode-Ray Tube

If we apply an alternating voltage to the Y-plate of a CRT, the electron beam will oscillate and the screen will present a luminous vertical bar (Fig. 16-10*a*). Its length is proportional to twice the peak value of the applied voltage, $2V_m$. If we know the deflection sensitivity of our CRT and have measured the deflection y, we can find V_m by the equation

$$V_m = y/2S_Y \qquad (16\text{-}5)$$

For example, if $S_Y = 0.4$ mm V^{-1} and $y = 20$ mm, we have

$$V_m = 20/(2 \times 0.4) = 25 \text{ V}$$

If the sensitivity is not known, it can be determined as follows: Apply a known alternating voltage to the deflection plates and measure the length of the luminous bar. The voltage can be drawn from an a. c. socket outlet and measured with a voltmeter. It should be remembered that the voltmeter will read the rms value of voltage which can be converted to its peak (or amplitude) value on multiplying it by 1.4.

As is seen, a CRT may be used as a peak-reading (or crest) voltmeter. Among its advantages are a high input resistance and the ability to make measurements at very high frequencies.

The above technique may be used to measure the peak value of nonsinusoidal voltages and also the amplitudes of the positive and negative half-cycles of voltage when they are not equal in magnitude. The procedure is as follows. Note the position of the light spot when the unknown voltage is not yet applied. Apply the unknown voltage, and measure the distance between the initial position of the light spot and the ends of the luminous bar, y_1 and y_2 (Fig. 16-10*b*). Then the amplitude of each half-cycle will be given by

$$V_{m1} = y_1/S_Y$$

and (16-6)

$$V_{m2} = y_2/S_Y$$

For visual observation of alternating voltages, the unknown voltage is applied to the Y-plates while the X-plates are fed what is known as the *sweep* or *timebase voltage*, V_{tb}, which has a sawtooth waveform (Fig. 16-11) and is generated by an oscillator commonly called the *sweep generator* or the *timebase generator* (or *circuit*). Owing to this timebase or sweep voltage, the electron beam moves horizontally across the screen at a constant rate for a time t_1 while the voltage is rising. This is the forward motion of the sweep, or the active phase of the *sweep scan*. During the next time interval, t_2, the voltage abruptly decreases, and the beam rapidly moves backwards. This is the *retrace motion* (or, simply, the *retrace*) of the sweep, also known as the *flyback*. These two events, forward sweep and flyback, repeat themselves at the frequency of the timebase voltage.

When the unknown voltage is not applied, the screen displays a luminous horizontal bar which is in effect the time axis. If we apply the unknown alternating voltage to the Y-plates, the spot on the screen will simultaneously oscillate up and down the screen and repeatedly sweep forward and fly back horizontally. The result will be a luminous pattern which represents the waveform of the unknown voltage. Figure 16-12 shows the waveform of a sinusoidal voltage, but visual observation of other waveforms is likewise possible.

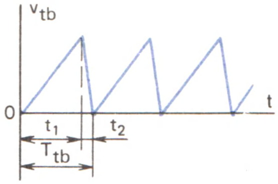


Fig. 16-11

Sawtooth voltage for a linear time-base

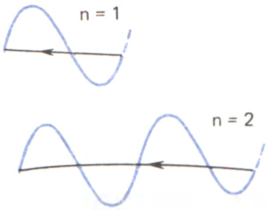


Fig. 16-12

Sinusoidal voltage waveforms at multiple frequencies

For the pattern displayed on the screen to appear stationary, the period of the timebase voltage, T_{tb}, must be equal to or be a whole multiple of that of the unknown voltage, T:

$$T_{tb} = nT \qquad (16\text{-}7)$$

where n is an integer (1, 2, 3, etc.).

Accordingly, the timebase frequency f_{tb} must be a submultiple of that of the unknown voltage:

$$f_{tb} = f/n \qquad (16\text{-}8)$$

Then the screen will be able to display a whole number of cycles of the unknown voltage during the time T_{tb}, and at the end of flyback the light spot will find itself at the same point where it started during the previous scan. The figure shows the waveform for $n = 1$ or $T_{tb} = T$ and also for $n = 2$ and $T_{tb} = 2T$. The flyback (or retrace) time, t_2, should preferably be made as short as practicable because some of the waveform is not displayed, that is lost, during flyback (the dashed portion of the pattern). Also, a shorter t_2 means a faster flyback, and so the retrace is less visible on the screen. If we wish to observe at least one cycle of the unknown signal, we should set n equal to at least 2. The value of n is chosen by adjusting the frequency of the sweep generator. If n is a non-integer, the waveform pattern will not remain stationary and several waveforms instead of one will be seen, which is an obvious inconvenience. Figure 16-13 shows the waveforms of a sinusoidal voltage for $n =$

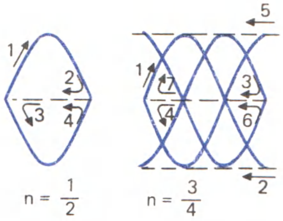


Fig. 16-13

Sinusoidal voltage waveforms at submultiple frequencies

= 1/2 and n = 3/4. For simplicity, it is assumed that the flyback time, t_2, is zero. The arrows and numbers in the figure indicate the sequence in which the spot moves on the screen.

The chosen number n will usually remain constant for a limited time only because the frequency of the timebase is anything but constant, and that of the unknown voltage may vary as well. To hold the number n constant for a long time, it is customary to synchronize the timebase with the unknown signal. Synchronization consists in that the unknown voltage is applied to the timebase and causes it to generate a sawtooth voltage at a frequency which is $1/n$ of that of the unknown signal.

The voltage to be observed is usually applied to the deflection plates via d.c. blocking capacitors (see Fig. 16-2). This prevents any d.c. component from reaching the plates, and the observer can only see the a.c. component. The time axis (the zero line) of this component is the horizontal trace remaining on the screen when the unknown voltage is no longer fed to the deflection plates. If it is desired to obtain a waveform of the entire signal, including its d.c. component, it must be appplied to the plates directly rather than via capacitors.

When the objective is to watch the waveform of a current, a resistor R is placed in the circuit. The voltage that develops across it is proportional to the unknown current and is applied to the Y-plates. Since the deflection sensitivity of the CRT is known, it is an easy matter then to find this voltage. On dividing it by the value of R, one obtains the unknown current. For the resistor not to affect the waveform or magnitude of the unknown current, its value should be relatively low. If the resultant voltage appears then to be insufficient, it must be fed via an amplifier with a known gain.

16-7 Image Distortion in Electrostatic CRTs

With electrostatic CRTs, waveform distortion is usually observed with an unbalanced connection of the deflection plates, that is, when one plate in each pair is connected to the second anode (see Fig. 16-2). Let, in this connection, an alternating voltage of peak value V_m be applied to the Y-plates. One of the plates will then be at zero potential with respect to the chassis ground, and the potential at the other will vary between $+V_m$ and $-V_m$ (Fig. 16-14*a*). The potentials at the various points in the space between the plates will vary in a similar fashion. During the positive half-cycles of voltage electrons move through points at potentials in excess of V_{a2}. As a result, they gain in velocity but the sensitivity of the tube is reduced. During the negative half-cycles electrons are slowed down because the potential at points between the plates are below V_{a2}, and the tube sensitivity is increased. As a result, the deflection y_1 during the positive half-cycles is smaller than the deflection y_2 during the negative half-cycles. The resultant waveform is nonsinusoidal which is another way of saying that nonlinear distortion has taken place.

With a balanced connection, none of the plates is connected directly to the chassis ground and the second anode, and points of zero

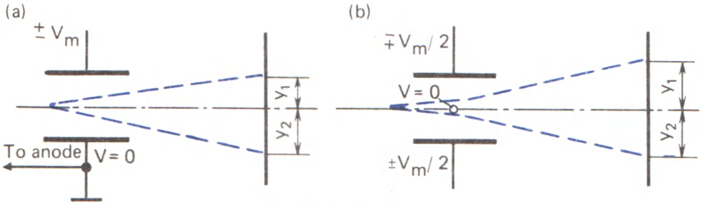


Fig. 16-14

Electron beam deflection in the case of (*a*) unbalanced and (*b*) balanced connection of the deflection plates potential lie midway between the plates (Fig. 16-14*b*). At any instant, the plates are at potentials which are of the same magnitude but opposite in sign. When the potential at one plate is $\pm 0.5V_m$, that at the other is $\pm 0.5V_m$. The deflection of the beam towards any plate occurs under identical conditions, and so $y_1 = y_2$. Figure 16-15 shows one of the likely arrangements for a balanced connection of the deflection plates. The direct voltage used to position the light spot initially (*positioning control*) is picked off twin resistors R_6 and R_6'. Their contact arms are moved by the same amount with a common knob so that the potentials at the deflection plates are changed by the same amount but in opposite senses.

The balanced connection of the deflection plates minimizes some other unpleasant consequences. For example, the light spot is less defocused as it moves towards the edge of the screen with a balanced connection of the plates.

An unbalanced connection of the plates more distant from the electron gun results in what is known as *keystone distortion*. It arises due to the presence of a field on the way of the electrons moving from one pair of plates to the other. Let, for example, the Y-plates located closer to the screen be connected in any arbitrary way and fed an alternating voltage, and also let the X-plates connected in an unbalanced manner be held at zero potential. The screen will then display a luminous vertical bar (at *1* in Fig. 16-16).

If we apply a positive potential to the X-plate not connected to chassis ground, the bar will be shifted towards that plate (at *2* in the same figure), but it will become shorter because an additional field has been set up between the positively charged X-plate and the Y-plates. This field somewhat bends the travel paths of the electrons and reduces their displacement caused by the voltage at the Y-plates. When a negative potential is applied to the same X-plate, the electrons emerging from the Y-plates will be acted upon by an additional retarding field which will somewhat increase their displacement; the luminuous bar on the screen will be shifted to the left and become longer (at *3* in the same figure). The three bars form a trapezoid instead of a rectangular shape. This form of distortion is called 'keystone' because the shape of the pattern displayed on the screen looks like the stone laid at the summit of an arch in order to lock the whole structure together. Keystone distortion can be controlled by placing shields between the X- and Y-plates or by shaping the plates more distant from the electron gun in some special manner.

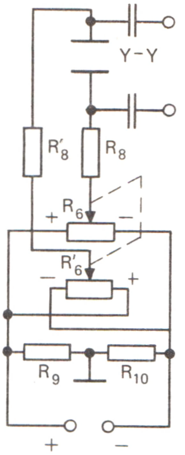


Fig. 16-15

An example of a balanced connection of the deflection plates

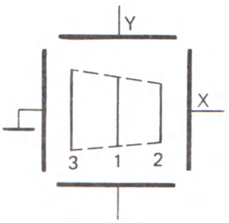


Fig. 16-16

Keystone distortion

Current practice is to use a balanced connection for the deflection plates because it minimizes many forms of image distortion. An unbalanced connection may be used only when the beam is to be deflected in one side only.

16-8 The Magnetic CRTs

Magnetically focused and magnetically deflected CRTs are most commonly used as TV picture tubes (*kinescopes*) and as *indicators* (*displays*) in radar. Because the focusing and deflecting coils are placed outside the CRT envelope, magnetic tubes are simpler in design than their electrostatic counterparts (Fig. 16-17). The electron gun of a magnetic CRT has

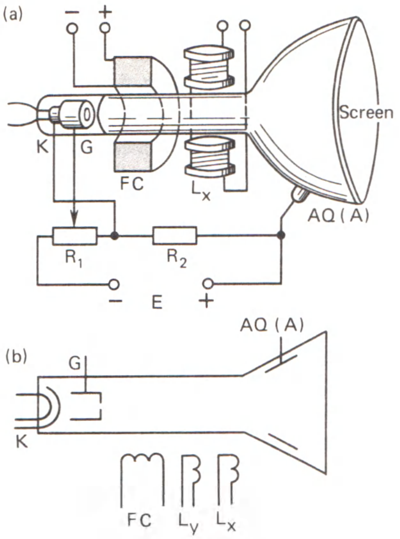


Fig. 16-17

Magnetic CRT: structure and graphical (circuit) symbol

a cathode, a control grid, and an anode. Sometimes the job of the anode is done by a conductive coating applied to the inside of the envelope as in Fig. 16-17*a*. Some magnetic tubes have a screen grid which is placed between the anode and the control grid and is held at a positive potential of several hundred volts. The electron gun is energized in the same manner as it is in electrostatic CRTs, but there is no need in anode voltage control for focusing purposes.

The diverging stream of electrons leaving the gun enters the magnetic field set up by a focusing coil, *FC*, which is energized with direct current. The figure shows a sectional view of the coil. Magnetic focusing may be accomplished with a long coil or a short coil. When a long focusing (or focus) coil is used, the electrons travel through a uniform magnetic field inside the long coil (Fig. 16-18) and their movement is along helical or screw lines. If the electrons leave point *B* on the coil axis, they will cross the axis each time around, that is, they will be focused at points B_1, B_2, etc. This is illustrated by projections of the travel paths onto a plane at right angles to the coil axis. They are circles emerging from point *B* and going back to the same point. (The figure shows the paths of only two electrons.)

Long-coil focusing is employed in some special-purpose devices. CRTs use short focus coils which act as thin magnetic lenses (Fig. 16-19). The motion of electrons in the field of a thin lens forms a complicated pattern, so we will consider it only approximately. To begin with, let us divide the field into two parts (*I* and *II*) by passing a plane through the middle of the coil at right angles to its axis. On either side of the plane, the magnetic induction (the magnetic flux density) decreases along the coil axis. When a diverging stream of electrons leaving point *B* enters part *I* of the field, the travel paths of the electrons are curved. In a uniform field, they would have been helices or spirals. In our case, however, they are far more complex lines owing to the fact that the field is anything but uniform.

In part *I* of the magnetic field the magnetic induction builds up. Therefore, the electron paths are curved more, and their curvature is a maximum at the boundary between regions *I* and *II*. Past that point, the magnetic induction decreases, and the travel paths of the electrons are curved less. On leaving the field, the electrons keep on moving by inertia – now they travel in straight lines which cut the tube axis at point B_1*. As is seen, the electrons travel along complicated spatial curves which may arbitra-

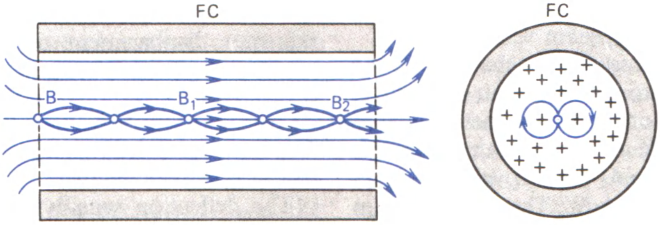


Fig. 16-18

Long-coil focusing

* In some texts, the electrons are erroneously shown moving in helices even outside the field.

rily be called helices or spirals of a variable radius. For better insight into how an electron moves in such a case, Fig. 16-19 shows projections of the trajectory on three mutually perpendicular planes. Because the electrons travel at a high velocity, these trajectories are only small fractions of one revolution of a spiral.

It is customary to enclose the focus coil of a magnetic CRT in a magnetically soft steel shield (Fig. 16-20). This arrangement enhances the magnetic flux density and improves the focusing action of the coil.

The mmf that the focus coil must supply for proper focusing is approximately given by

$$F_M = Iw \approx 240\,(V_a d/l)^{1/2} \qquad (16\text{-}9)$$

where d = mean coil diameter, cm
l = distance from the coil to the screen, cm
V_a = anode voltage, kV
w = number of coil turns
I = coil current, A

Ordinarily, the focus coil has several hundred or even thousand turns. For example, at $I = 0.1$ A, $d = 6$ cm, $l = 18$ cm and $V_a = 3$ kV, its mmf will be $F_M = 240$ A, and it will have $w = 240/0.1 = 2400$ turns.

A well-designed shielded coil needs markedly fewer turns. A further improvement is assured through control of the coil current with a potentiometer. The direction of current in the focus coil is of no importance. Sometimes, the coil is replaced with a ring-shaped permanent magnet, and the focus is controlled by moving the magnet along the tube neck or by moving a magnetic shunt into which a fraction of the total magnetic flux is divided.

The magnetic deflection of the electron beam uses two pairs of coils positioned at right angles to each other. For simplicity, Fig. 16-17*a* shows only one pair with a vertically directed field – these are the X-coils of the tube. The other pair, or the Y-coils, produces a horizontally directed field which deflects the beam vertically.

If we take it approximately that inside the tube the field set up by each pair of coils is uniform, the electrons may be thought of as moving along an arc of a circle with point O as centre inside the field, and in straight lines outside the field (Fig. 16-21). The electron beam is caused to deflect through an angle α and the light spot on the screen is displaced through a distance y. The vertical deflection sensitivity of a magnetic CRT may be defined as the ratio of the spot displacement on the screen to the magnetizing force that causes this displacement:

$$S_Y = y/F_Y = y/I_Y w_Y \qquad (16\text{-}10)$$

The horizontal deflection sensitivity, S_X, is defined in a similar manner.

The deflection sensitivity of state-of-the-art magnetic CRTs does not exceed a few tenths of a millimetre per ampere. It depends on the design of the tube and deflection coils and its

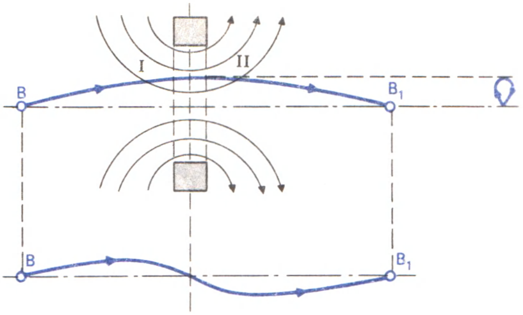


Fig. 16-19
Short-coil focusing

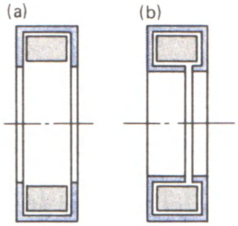


Fig. 16-20
Shielded focusing coils with (*a*) a wide slot and (*b*) a narrow slot

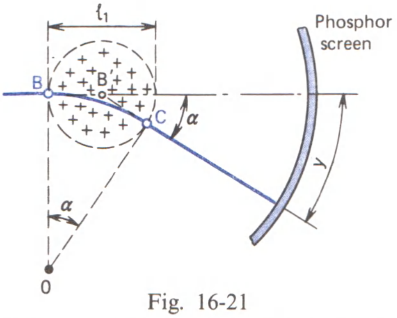


Fig. 16-21
Beam deflection in the magnetic field of the deflection coils

operating conditions. This relation takes the form

$$S_Y = \gamma l / V_a^{1/2} \qquad (16\text{-}11)$$

where l is the distance of the coil axis to the screen (in mm) and γ is a coefficient taking care of the coil design, usually equal to 0.1-0.2 $\mathrm{V}^{1/2}\,\mathrm{A}^{-1}$.

For example, if $\gamma = 0.15$, $l = 200$ mm and $V_a = 2.5$ kV, the vertical deflection sensitivity will be

$$S_Y = 0.15 \times 200 \times 2500^{-1/2} = 0.6 \text{ mm A}^{-1}$$

The value of γ for a given type of deflection coils can be found by experiment. For this purpose, one first finds S_Y by Eq. (16-10), then measures l and V_a, and finally computes γ by Eq. (16-11).

The sensitivity of magnetic CRTs is less dependent on the anode voltage (V_a is the radicand) than that of electrostatic tubes. However, the two types of tubes ought not to be compared in terms of this constant because different units are used.

One way to boost the magnetic field in a CRT is to use closed magnetic cores fabricated from magnetically soft steel or other ferromagnetic materials. At high frequencies, no cores are usually employed, and the coils are given some special shape. They enclose the tube and produce a more uniform field. To minimize magnetic flux leakage, the coils are enclosed in ferromagnetic shields.

In the past, magnetic focusing would usually outperform electrostatic focusing, but in state-of-the-art CRTs electrostatic focusing is at least as good as is magnetic focusing. With magnetic focusing, the CRT is simpler in design because the focusing coil or magnet is exterior to the envelope rather than placed in the vacuum inside the tube. An advantage of electrostatic CRTs is the economy of operation because they need no power to produce a current in a focusing coil. In contrast, magnetic deflection calls for a fairly large supply source to energize the deflecting coils. On the other hand, magnetic deflection simplifies the tube structure and permits the beam to be deflected through a larger angle. As a result, even a large-screen CRT may be made a good deal shorter. With magnetic deflection, one need not be worried over the image distortion described in Sec. 16-7. It is to be noted that the inductance of the deflecting coils adds to the inertia (time lag) of the deflection process, and so a magnetic deflection system cannot show a sufficiently good performance at very high frequencies. The input impedance of the deflection coils is small at low frequencies, but it is further decreased at high frequencies because of the self-capacitance of the coils. The input impedance of electrostatic CRTs remains sufficiently high even at high frequencies.

16-9 The Fluorescent Screen

As has been noted, the screen of a CRT is given a coat of a material which fluoresces when bombarded by electrons. Such a material is called a *phosphor*. To obtain the desired brightness (intensity), colour and persistence, it is usual to add activators to the phosphor material. These activators are most commonly silver, manganese, or copper. The long persistence of radar CRTs is assured through the use of copper as the activator. A silver-activated phosphor used in TV picture tubes (kinescopes) produces a medium persistence.

The most commonly used phosphors have the following properties. Zinc oxide has a short persistence which is important for oscilloscopes and produces a violet or green emission colour. Various mixtures of zinc sulphide and cadmium sulphide produce a bright emission of any colour, notably white, with a persistence of a fraction of a microsecond to several minutes. Phosphors for visual observation are prepared from zinc silicate which may be artificially prepared or occurring naturally (as the mineral willemite) with some manganese added to act as the activator. They produce a short persistence and their emission colour ranges from green to yellow-orange. A blue-violet emission colour with a short persistence is produced by screens with their phosphor prepared from the tungstates of barium, calcium, magnesium, cadmium, zinc or strontium.

The intensity (or brightness) of the light emitted by a phosphor screen is approximately proportional to the square of the potential difference between the screen and cathode. In other words, it increases with increasing velocity of the electrons in the beam. There is a certain minimal electron energy that is necessary for the emission of light to occur. It ranges from tens to hundreds of electron-volts. At lower energies, electrons are unable to penetrate the lattice of the phosphor crystals. At energies of several kiloelectron-volts the penetration depth does not exceed 1 μm. At low beam-current values the brightness of the spot is proportional to the current density, but it ceases to increase after the latter has reached a certain value (the saturation effect).

The efficiency of a phosphor, defined as the ratio of the energy carried by the visible emission to the total energy of the impinging electrons, does not exceed a few per cent. The greater part of the beam energy is wasted as heat dissipated at the screen, expended to knock out secondary electrons, and radiated as ultra-violet and X-rays.

The performance of a phosphor screen can conveniently be stated in terms of its *luminous efficiency* (or *light output*) defined as the luminous flux per watt of beam power. The luminous efficiency of a phosphor is a maximum when the temperature of the phosphor remains in the range 0 to 60-80°C. Any further rise in screen temperature causes a fall in the luminous efficiency; at 400°C there is no light emission at all.

It takes a certain time for a phosphor screen to come up to its full emission after electron bombardment has begun. This is known as the *excitation time* of a phosphor. After the bombardment has ceased, it again takes some time for the screen emission to decay. This is known as the *persistence* of a phosphor. At first, the emission decays at a high rate, then it falls off exponentially. This exponential decay characteristic is described by the term '*time constant*', and it is the time required for the intensity of emission to fall to $1/e$ of its initial value (at the time when the excitation ceases). In Soviet practice, the time constant is often referred to as the time for the intensity to decrease to 1% of the initial value. The persistence of a phosphor may be very short (less than 10^{-5} s), short (from 10^{-5} to 0.01 s), medium (from 0.01 to 0.1 s), long (from 0.1 to 16 s), and very long (over 16 s).

An important factor in the operation of a phosphor is secondary electron emission. The secondary emission ratio σ depends on the primary-electron energy which is in turn dependent on the screen potential V_s with respect to the cathode. It is a maximum when the impinging electrons have an energy of several hundred electron-volts, following which it decreases (Fig. 16-22). A screen will produce a steady glow if its potential remains constant. This can happen only when the number of electrons arriving at the screen is equal to that of secondary electrons ejected from the screen. This is a steady state condition. Obviously, phosphors with $\sigma < 1$ are not suited for screens. A good phosphor should have $\sigma > 1$.

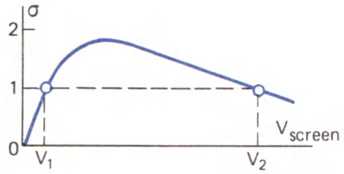


Fig. 16-22

Secondary emission ratio of a phosphor as a function of primary-electron energy

So long as the screen potential stays at below V_1, no emission can take place because at $\sigma < 1$ the screen potential will further be brought down by impinging electrons. If the screen potential is anywhere between V_1 and V_2, then σ will be greater than unity, and the screen will have a steady-state potential by several volts higher than that of the second anode and of the Aquadag coating connected to it. In the circumstances, the secondary electrons will see a retarding field which will drive some of them back to the screen. The remaining secondary electrons will be able to reach the Aquadag coating because they have higher initial velocities. The current constituted by the secondary electrons is equal to the beam current. Because the Aquadag coating and the screen are usually at high potentials with respect to the cathode, we may deem that, on neglecting a difference of a few volts, they are at the same potential.

If, on the other hand, the initial screen potential V_s is higher than V_2, the impinging electrons will bring it down to a value very close to the second-anode potential V_{a2} because the number of arriving primary electrons is equal to that of ejected secondary electrons. The potential V_2 is the highest attainable one for a given phosphor; it is called the *critical potential* for the phosphor. It differs from one phosphor to another and ranges anywhere between 5 and 35 kV. The value of critical potential is a vital factor for CRTs. The higher its value, the greater the velocity that the electrons can attain in the beam, and the brighter the image seen on the screen.

Obviously, it would not pay to set V_{a2} above the critical potential V_2 because the velocity of electrons as they strike the screen is decided by V_s and not by V_{a2}. For example, if $V_{a2} = 10$ kV and $V_s = 6$ kV, electrons will leave the second anode with an energy of about 10 keV, but on their way from the anode through the retarding field to the screen they will lose 4 keV and will hit the screen with an energy of 6 keV. But the same situation would exist at $V_{a2} = 6$ kV.

Electron bombardment gradually reduces the light output of the phosphor screen, but the original value of light output is restored if the screen is allowed to 'rest'. When a CRT has been in operation non-stop for a long time, a nonreversible degradation of light output occurs, known as the phosphor burn. The areas on the screen that have been bombarded by electrons most of all grow darker in proportion to the beam power. An increase in the beam current density aggravates the phosphor burn more than an increase in the electron velocity. Therefore, it is more advantageous to use a higher anode voltage at a lower beam current. It is to be recalled that a rise in V_{a2} improves focusing as well.

It is desirable that the image presented on the screen have a just-sufficient brightness. Care must be taken to avoid a very bright but stationary spot appearing on the screen as this would lead to the phosphor burn. If it carries enough power, a stationary electron beam may well melt the glass.

The phosphor deteriorates when bombarded by the negative ions that are emitted by an oxide cathode along with electrons. Larger in mass, the ions retain a nearly straight-line travel path despite the action of magnetic field. Therefore, in magnetic CRTs the ions travel as an unfocused stream and bombard one and the same central part of the screen all the time, thus producing a dark *ion spot*. To avoid it, resort is made to electron guns that incorporate an *ion trap* (see Sec. 16-10).

Within the ion spot, the surface layer of the phosphor is completely burned. If we raise the anode voltage, the electrons will move deeper into the phosphor and elicit an intensive fluorescence. This can remove the ion spot completely or in part for some time. Of course, the anode voltage ought not to be raised above its absolute maximum (or safe) rating. In electrostatic CRTs, ions are focused and deflected similarly to electrons, so no ion spot is formed on the screen. With time, however, the emissivity of the screen deteriorates, the critical voltage is brought down, and this leads to a lower brightness.

The performance of the screen can be improved by giving it a coat of aluminium 0.1-0.2 μm thick on the side facing the beam. The aluminium coating is connected to the conductive Aquadag coating. Aluminized screens offer a number of advantages. For one thing, secondary emission from the phosphor is no longer needed. The conductive aluminized backing provides an easy path for the electrons emitted by the screen to enter the second-anode circuit. Therefore, the critical voltage at the screen may be a good deal higher than for an unaluminized screen. In consequence, the electrons inside the CRT can travel at higher velocities, and this leads to a greater brightness. The brightness is enhanced still more by the reflection of light from the aluminium backing. The slower ions are unable to break through the aluminium backing, and no ion spot can be formed. In contrast, the faster electrons readily move through the aluminium film into the phosphor, although they lose some of their energy in the process.

Aluminized phosphor screens are used in CRTs operated at high anode voltages. At low anode voltages it is not warranted to use an aluminized phosphor screen because too much of the electrons' energy would be lost in breaking through the aluminium.

The image displayed on the screen should preferably be maximally sharp, crisp, and of proper contrast. Unfortunately, the objective cannot always be achieved owing to several factors. Contrast may be impaired by extraneous light falling on the screen, unless the picture is watched in a darkened room. The reduction in contrast leads in turn to the appearance of a *halo*, a ring of light around the luminous spot on the screen, thereby impairing sharpness. Sometimes, two or more halos may be seen. How a halo is brought about is explained in Fig. 16-23. The bulk of the light rays emitted by the spot passes outside through the glass, but the rays incident at large angles to the outer surface of the glass experience a total internal reflection, go back to the phosphor and spread in it, giving rise to a halo. Some of these rays may again experience a total internal reflection and give rise to a second halo, etc.

Contrast may markedly be reduced owing to the reflection of light from the inside of the flared part of the envelope (Fig. 16-24*a*). This can be avoided by giving the tube envelope some special shape (Fig. 16-24*b* and *c*).

The rays emitted by the light spot may directly illuminate the screen owing to its curvature (Fig. 16-25). This drawback would be nonexistent in a CRT with a flat screen. However, the faceplate of large CRTs has to be made slightly convex so that the tube envelope could stand up to the higher atmospheric pressure. This illumination by light rays from the spot does not occur in CRTs with an aluminized phosphor screen because the aluminium does not let light pass inside the tube. Scattered electrons produced by secondary or electrostatic emission from the electrodes may also cause a faint fluorescence of the phosphor.

16-10 Basic Types of CRTs for Oscilloscopes and TV

Oscilloscopes mainly use electrostatic CRTs. Radar and sonar indicators ordinarily employ magnetically deflected CRTs in which the beam deflection may be magnetic or electrostatic. In radar and sonar indicators, target presentation is usually by *intensity modulation*: the target echo is applied to the control grid and turns on ('unblanks') the tube. Owing to magnetic deflection, these tubes suffer less from nonlinear distortion and their beam is better focused even at large deflection angles. Sometimes, double-beam CRTs are used in which each tube encloses two electrode structures in a common envelope and permits visual observation of two events at the same time.

TV picture tubes, or kinescopes, are usually made with magnetic deflection while the focusing system may be magnetic or electrostatic. Magnetic deflection improves focus and enhances brightness because a markedly higher anode voltage may be used. Some kinescopes come enclosed in metal-glass envelopes.

Many kinescopes have ion traps (or *beam benders*) which prevent negative ions from reaching the screen and burning a spot in it. As a rule, ion traps operate by separating the ion stream from the electron stream with the aid of a magnetic field. One design of the ion trap is shown in Fig. 16-26. The axis of the cathode, control grid and screen grid makes an angle with the tube's axis, and the anode's axis makes a kink. On entering the anode, the stream of negative ions (the full lines) and of electrons (the dashed lines) find themselves in the transverse field of a permanent magnet (the shaded area). The heavier ions are not practically deflected by the magnetic field and reach the anode. The electron paths, however, are bent, and the electrons leave the anode through the hole in its end. The permanent magnet of the ion trap is placed externally to the tube, and its optimal position is found by trial and error.

State-of-the-art kinescopes have a rectangu-

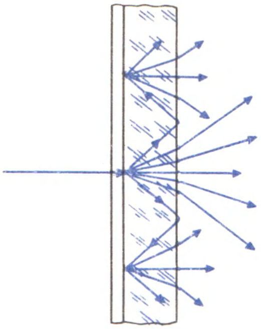


Fig. 16-23

The formation of a halo around the light spot

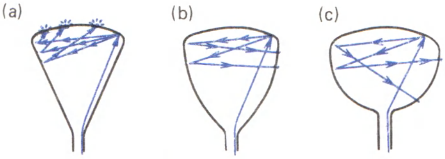


Fig. 16-24

Effect of the envelope shape on the reflection of light from its walls

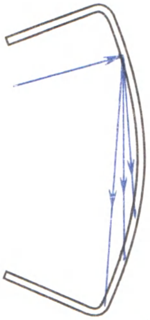


Fig. 16-25

Stray illumination of a spherical screen by rays from the light spot lar screen and a beam deflection angle of 110°. Because of the large deflection angle, they are shorter than the older 70-deg tubes.

16-11 Shaped-Beam Character Display Tubes

Recent years have seen a growing interest in *shaped-beam tubes* used to display letters, numbers, and symbols at rates compatible with readout from digital computers. The information displayed on the tube face may be viewed directly on a real-time basis or it may be recorded on film or other media for later analysis.

In addition to their use as character displays, shaped-beam CRTs can be used in what may be called the '*spot-writing*' mode which allows regular *raster scan* or *PPI displays* (such as used for radar or sonar ship-borne equipments) to be presented.

One representative of this class of CRTs is the *Charactron* which is the trade name that General Dynamics/Electronics (USA) uses for a specially designed cathode-ray tube capable of displaying alphanumerical characters and other special symbols directly on its screen.

The way the Charactron can be used in a net of radar stations is shown in Fig. 16-27. Several radars (or sonars) are connected to a computer which processes raw data about objects (or '*targets*' as they are usually called in radar and sonar parlance). The signals generated by the computer drive a device that controls the Charactron. The various targets detected by the radars (or sonars) are displayed on the Charactron screen as an array of letters, numbers, etc. Several targets are thus displayed at the same time, with the arrays located on the screen according to the targets' coordinates and movements (Fig. 16-28). In this way, a single Charactron can perform the situation surveillance function earlier performed by several individual radar scopes.

The relative location of the elements in one type of tube is shown in Fig. 16-29. The electron beam shown as a dashed line is generated by an electron gun, *EG*. The two pairs of deflection plates, called the *selection plates* (*SP*), guide the beam onto a *matrix* or *mask*, *M*, a metal plate in which openings are cut to the shape of various characters. There may be as many as several tens of openings, with their size not exceeding a few tenths of a millimetre, which is somewhat smaller than the beam diameter. Selection of the desired character is accomplished by means of a code from a computer. The code is fed into selection decoders which convert the input information into X and Y deflection voltages. These voltages, after passing through amplifiers, are applied to the deflection plates and cause the electron beam to deflect to one of the character positions in the matrix.

On passing through the matrix, the beam is deflected away from the tube axis. If the beam were not returned to the optical axis, a nonuniform display would result on the screen. To overcome this, two more elements are added. They are the *convergence coil, CC*, and the *correction plates, CP*. The magnetic field of the

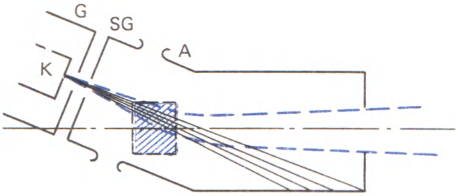


Fig. 16-26
Ion trap

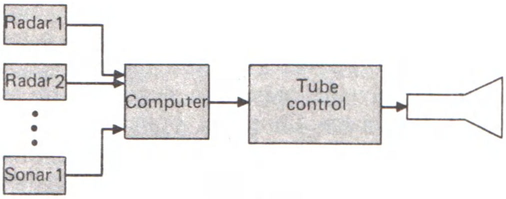


Fig. 16-27
Block diagram of a radar and a sonar equipment using a Charactron-type CRT

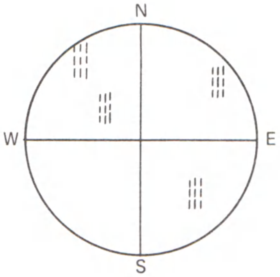


Fig. 16-28
Pip arrays on the screen of a Charactron

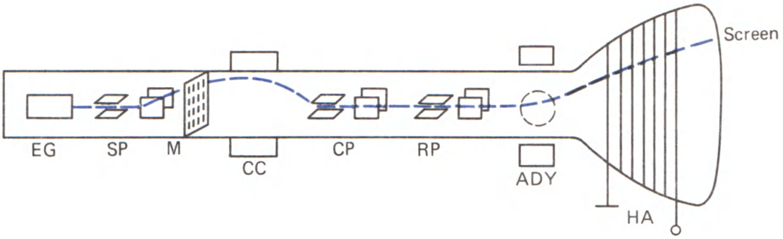


Fig. 16-29
Structure of the Charactron

convergence coil re-directs the electron beam towards the optical axis, and the correction plates supply the correction necessary to redirect the beam along the optical axis and through the *reference plates, RP*. Of course, the voltages across the *CC, CP* and *RP* are matched with that across the *SP*. The convergence coil has additional windings to make up for the tilt of the characters due to the action of the magnetic field set up by the main coil.

Deflection of the shaped beam to the desired position on the tube face is accomplished by an *address deflection yoke, ADY.* The electron gun of the Charactron operates at a relatively low voltage, so the electrons in the beam travel at a moderate velocity. For this reason, the beam can be deflected with a moderately high voltage and current, and this simplifies the tube design. The brightness of the displayed characters is enhanced by post-deflection acceleration. The post-deflection accelerating anode is made in the form of a high-resistance conducting helical layer and is therefore called the *helical accelerator, HA.* The post-deflection accelerating voltage rises from one turn of the helix to the next and this minimizes character distortion on the screen.

There are other types of shaped-beam character display tubes in which the deflection plates are replaced with deflection coils or the deflection coils are replaced with deflection plates and which use some other elements.

The screen diameter of a shaped-beam character display tube may be several tens of centimetres. The characters as they appear on the screen may be as large as 2.5-3.5 mm. To avoid flickering, the displayed array is projected at a rate of 15-20 times per second.

The rates at which state-of-the-art shaped-beam character display tubes can operate along with the associated computer may be as high as tens of thousands of characters per second.

Chapter Seventeen

Gas-Discharge and Indicator Tubes

17-1 Gaseous Discharges

Gas-discharge (*gas-filled* or, simply, *gas*) *tubes* is the name used to describe partially evacuated electron tubes containing a small amount of gas. Ionization of this gas is responsible for the current flow in such tubes. In fact, this ionization is only a part of the more general process known as a *gaseous discharge*. When a discharge occurs in a gas (or, which about the same, a vapour), several fundamental events take place there.

Excitation. When an electron hits an atom, one of the latter's electrons is driven to an orbit more distant from the atomic nucleus, that is, to a higher energy level. The target atom is then said to be in an excited state. This excited state lasts usually for 10^{-7}-10^{-9} s, following which the electron falls back to its normal orbit (or the normal energy level). In doing so, it gives up the energy it has received from the incident electron as a packet of radiant energy. The emission of radiation will be accompanied by a glow if the radiant energy falls within the visible portion of the electromagnetic spectrum.

For an atom to be excited, the impinging electron must possess a certain energy called the *excitation energy*.

Ionization. When an impinging electron carries more energy than is necessary for excitation, the target atom loses an electron and turns into a positive ion, with two free electrons existing in the interatomic space. If these two electrons acquire enough energy while moving through an accelerating field, each will be able to turn into an ion, or ionize, one more atom. This will produce four free electrons and three ions. Each of these electrons may again ionize one more atom, and so on. In this way, the number of free electrons and positive ions grows in a cumulative, avalanche-like manner.

Ionization may also occur in a stepped fashion. Hit by an electron, an atom moves into an excited state and is further ionized by a second electron before it has time to fall back to its normal state.

The increase in the number of charged particles in a gas owing to ionization is called *gas amplification*.

The excitation and ionization energies (in electron-volts) for some gases are given below.

	W_{exc}	W_{ion}
Argon	11.6	15.5
Helium	20.8	24.5
Hydrogen	11.1	13.5
Krypton	10.4	14.0
Neon	16.6	21.5
Xenon	8.4	12.1

Recombination. The ionization of a gas is accompanied by a reverse process – the simultaneous elimination of both an electron and an ion. The positive ions and the electrons in a gas are in a haphazard (thermal) motion. When an ion and an electron approach each other closely enough, they may combine (or, rather, recombine) to form a neutral atom. This is promoted by the mutual attraction of unlike charged particles. Quite logically, this process is termed *recombination*. The neutral atom emerging from recombination may again be ionized, and the resultant positive ion and electron may again recombine, etc.

Recombination leads to a reduction in the number of charged particles, that is, the *de-ionization* of the gas. In fact, there may be either an increase or a decrease in the number of charged particles at any given instant, according as ionization or recombination gets the upper hand. In a steady state the number of electrons (or ions) produced every second due to ionization is equal to the number of neutral atoms produced by recombination. Notably, ionization prevails over recombination when an electric discharge occurs in the gas. As the discharge decays, recombination prevails over ionization. When the discharge ceases completely, ionization occurs no more, and recombination restores the gas atoms to their neutral state.

Since some energy is expended to bring about ionization, the positive ion and the electron it produces have between them a greater amount of energy than a neutral atom. This is the reason why recombination is accompanied by the release of radiant energy. As a rule, there is a visible glow of the gas.

It takes some time for recombination to take place, and so an ionized gas will usually be de-ionized in a matter of 10^{-5}-10^{-3} s. Because of this, gas tubes show a greater time lag, or inertia, than do vacuum tubes, and they are therefore unable to operate at high frequencies. The basic cause of this time lag is the low rate of de-ionization (a discharge occurs in 10^{-7}-10^{-6} s – thus a gas is electrified a good deal faster).

17-2 Forms of Gaseous Discharge

Above all, gaseous discharges may be *self-sustained* (or *self-maintained*) and *nonself-sustained* (or *nonself-maintained*). A self-sustained discharge is maintained solely by the action of voltage. A nonself-sustained discharge can only exist on the proviso that the action of a voltage is supplemented by some other extraneous factors. They may be light, ionizing radiation, thermionic emission from a hot cathode, etc.

Both types of discharge can exist in several forms. Some of them are briefly described below.

Dark (or *silent*) *discharge* is a nonself-maintained discharge. It is characterized by a current density of several microamperes per square centimetre and by a space charge of a very low density. The field set up by the applied voltage in the case of a dark discharge does not practically depend on the density of the space charge so the effect of the latter may be neglected. There is usually no glow of the ionized gas. This form of discharge is not utilized in gas-filled tubes for electronic circuits, but it precedes other forms of discharge.

Glow discharge is a self-maintained discharge. As its name implies, this discharge has a soft glow of light not unlike that of smouldering charcoal. The current density in a glow discharge is units or tens of milliamperes per square centimetre and there is a space charge markedly affecting the electric field between the electrodes. For a glow discharge to take place the voltage must be tens or even hundreds of volts. The discharge is maintained by the emission of electrons from the cathode bombarded by ions.

This type of discharge is mainly used in *voltage stabilizer* (voltage regulator or voltage reference) diodes, glow-discharge lamps, glow-discharge (cold-cathode) thyratrons, character-display tubes, and decade-counting tubes.

Arc discharge may be self-sustained and non-self-sustained. It involves current densities greatly in excess of those encountered in glow-discharge devices. Devices using a nonself-sustained arc discharge include gas-filled rectifier diodes and hot-cathode thyratrons. Mercury-arc tubes (of the excitron type) and ignitrons which have a mercury-pool cathode, and also triggered spark gaps utilize a self-maintained arc discharge.

In an arc-discharge device the current density may be several hundred amperes per square centimetre and the space charge strongly affects the events taking place in the gas. The arc discharge is sustained by thermionic emission from a hot solid cathode or electrostatic emission from a mercury-pool (liquid) cathode. During an arc discharge, nearly all of the voltage (10-20 V) is concentrated near the cathode. A low voltage drop in combination with a heavy current is typical of the arc discharge. It is accompanied by an intensive glow of the gas. An arc discharge may take place not only at a reduced pressure, but also at the normal or an elevated pressure, such as in cine projectors and searchlights.

Spark discharge is not unlike an arc discharge. It is a short-duration (pulse) electric discharge which can take place at a relatively high (say, atmospheric) pressure. As a rule, a spark is a series of pulse discharges following one another. This type of discharge is utilized in spark gaps used to make and break some circuits temporarily.

High-frequency discharges can occur in a gas under the action of an alternating electromagnetic field even if there are no current-conducting electrodes.

Corona discharge is a self-sustained discharge. It is utilized in gas-filled voltage regulator (VR) tubes. It takes place at relatively high pressures when at least one of the electrodes has a pointed tip or a small curvature or is made in the form of a fine wire. The field between the electrodes is then nonuniform, and the field intensity increases enormously at the pointed electrode. As a rule, this is the anode. The corona layer emits photons (quanta of radiant energy), and these ionize the gas at the boundary between the corona layer and the outer region, thus producing free electrons. The electrons move towards the anode, and excite (ionize) other atoms on their way there. In the outer region which remains dark, neither excitation nor ionization occurs because the field is of low intensity, and there is only a flow of particles charged similarly to the corona electrode, that is, positive ions.

Because a corona discharge affects only a part of the discharge space, it is regarded as an incomplete breakdown of the gas (a complete breakdown occurs in the case of a spark or arc discharge). As the voltage drop across the electrodes is increased, the current builds up, the corona layer expands, and the corona turns into a spark discharge if the pressure of the gas is considerable, or into a glow discharge if the gas pressure is low.

17-3 Glow Discharge

We will examine the glow discharge that takes place between two parallel-plane (flat) electrodes as shown in Fig. 17-1. In the absence of a discharge when there is no space charge, the field is uniform and the potential between the electrodes is distributed in a linear fashion (curve *1*). In a vacuum tube, thermionic emission gives rise to a negative space charge which sets up a potential barrier near the cathode (curve *2*). This barrier does not allow the anode current to rise high. In a gas-filled glow-discharge tube, a positive space charge is set up by a large number of positive ions. It causes the potential in the anode-cathode space to shift in the positive direction. The potential diagram 'flexes' downwards (curve *3*).

As is seen, the potential distribution in a gas-filled tube is such that nearly all of the anode voltage is applied to a thin layer of gas at the cathode. This is the *cathode fall of potential region* (*I*). A strong accelerating field is thus produced at the cathode. The anode is as it were moved closer to the cathode. The role of the anode is played by the positive ion cloud shadowing the cathode. As a result, the effect of the negative space charge is neutralized and there is no potential barrier at the cathode.

The other region of the glow discharge space (*II*) is characterized by a small voltage drop, and the field in this region is of low intensity. This is the *electron-ion plasma region*. Generally, a plasma is a heavily ionized gas in which the positive ions and the negative electrons are roughly equal in number. In a plasma, the haphazard (thermal) motion of particles prevails over their ordered motion. Still, the electrons move towards the anode and the ions towards the cathode.

The field forces acting on the electrons and ions are equal in magnitude but opposite in direction because the respective charges are equal in magnitude but opposite in sign (as will be recalled, the force exerted on a charge e is $F = eE$, where E is the field strength). However an ion has a mass which is thousands of times that of an electron. Even for the lightest gas, hydrogen, the mass of a positive ion is 1840 times the mass of an electron. Therefore, the ions are accelerated less and acquire relatively low velocities. In consequence, the current in a gas-filled tube is practically constituted by electron movement. The share of the ion current is very small and it may be neglected. The ions do a job of their own: they produce a positive space charge which substantially exceeds the negative space charge and eliminates the potential barrier at the cathode.

The cathode fall of potential region plays an important role. The ions entering this region from the plasma are accelerated. Impinging on the cathode at a high velocity, they knock electrons out of it. This process is essential for the discharge to be sustained. If the ions were too slow, no electrons would be emitted, and the discharge would cease. Emitted by the cathode, the electrons are likewise accelerated in the cathode fall of potential region and are capable of ionizing the gas atoms. Since electrons collide with gas atoms in various parts of the plasma region, ionization takes place throughout the gas volume. Of course, it is accompanied by recombination.

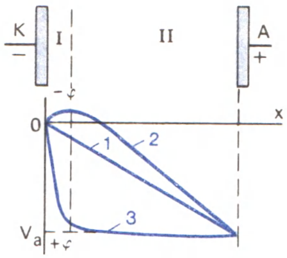


Fig. 17-1

Potential distribution between the electrodes (*1*) in the absence of an electric discharge, (*2*) in a tube, and (*3*) a gas tube in the case of a glow discharge

It is important to remember that only a small proportion of the ions produced in the plasma bring about electron emission from the cathode. Most ions recombine with electrons and fail to reach the cathode. When a glow discharge occurs, the number of ions hitting the cathode every second is such that they knock out as many electrons as were ejected during the previous second. These newly ejected electrons produce in the plasma as many ions as appeared there during the previous second, and again a certain proportion of these ions reaches the cathode to knock out of it the previous number of electrons per second. This chain of events repeats itself every second and sustains the glow discharge at a certain value of current.

The glow discharge has a soft glow of light at the cathode. This glow increases, expands and finally occupies all of the plasma with rising current.

A glow discharge can exist if the voltage drop across the diodes is not below a certain value. If the voltage drop is not high enough, the ions hitting the cathode will be unable to knock any electrons out of it. A nonself-sustained dark discharge turns into a self-sustained glow discharge at a certain breakdown voltage which in this case is the *firing, starting* or *striking potential*, V_{st}.

The striking potential V_{st} is a function of the gas used, its pressure, the material and separation of the electrodes. It is lower for an activated cathode. Figure 17-2 is a plot relating the striking potential V_{st} to the product of the gas pressure p and the electrode separation d. This plot is usually called the *glow-discharge characteristic*. The minimum value of striking potential, $V_{\text{st min}}$, corresponds to a value of the product pd which may arbitrarily be called optimal. In many gas-filled tubes, it is more advantageous to use other operating conditions.

The shape of the plot in Fig. 17-2 may be explained as follows. Let the electrode separation or spacing, d, be constant. Then at a very low pressure the occurrence of a discharge will be impeded because few electrons collide with atoms. As a result, few ions are produced, and they do not knock a sufficient number of electrons out of the cathode. The applied voltage has to be raised so that the ions could acquire a sufficiently high velocity and knock out a sufficient number of electrons. At a higher gas pressure, electrons collide with gas atoms too often and fail to come up to the velocity required for ionization to take place. As a result, too few ions are produced. An increase in the applied voltage raises the energy of the electrons, boosts the ionization, and leads to the occurrence of a glow discharge. As is seen, V_{st} has to be raised both when the gas pressure is low and high. At some medium pressure, on the other hand, a certain minimal value of V_{st} is sufficient.

When the gas pressure is held constant and the electrode spacing is very small, most electrons reach the anode without colliding with the gas atoms. Few ions then are produced and a higher voltage has to be applied so that they could knock a sufficient number of electrons out of the cathode. When the electrode spacing d is very large, the field strength is reduced. On their way to the cathode, electrons collide repeatedly with gas atoms and fail to acquire the energy required for ionization. The applied voltage has to be raised to a value such that between two consecutive collisions an electron could fall through a potential difference which is not less than the ionization potential. Thus, V_{st} has to be raised both when the electrode spacing d is small and large. At some medium value of d, a certain minimal value of V_{st} will suffice. Each gas has a characteristic of its own, similar to that shown in Fig. 17-2.

The voltage-current characteristic of the glow discharge is measured, using the set-up shown in Fig. 17-3. The dot on the graphic symbol of a gas-filled tube indicates that its envelope is filled with a gas. At one time, the tube symbol was shaded to indicate this fact.

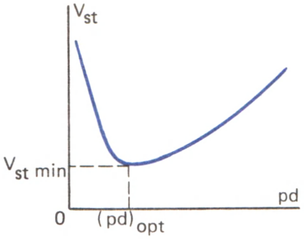


Fig. 17-2

Firing characteristic of a glow-discharge tube

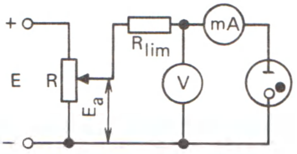


Fig. 17-3

Test set-up to measure the current-voltage characteristic of a gas tube

In a circuit, a gas-filled tube should be connected in series with a *current-limiting resistor*, R_{lim} of a proper value. If this resistor has too high a value (tens or even hundreds of megohms), a supply voltage of several hundred will produce a dark discharge because the current will not exceed several microamperes. At a substantially lower value of R_{lim}, glow discharge will take place, provided the supply voltage is not less than V_{st}.

A further decrease in the value of R_{lim} may lead to an arc discharge. This is an undesirable occurrence for glow-discharge tubes designed for a current of not more than several tens of milliamperes. When an arc discharge occurs, the current rises many times, and the tube may fail. If a gas-filled tube is connected to a high-voltage, low-resistance source directly, that is, without R_{lim}, this will likewise lead to an arc discharge. The current will mainly be limited by the internal resistance of the source because the resistance of a gas tube under conditions of an arc discharge is very small. The source will be short-circuited, the current will rapidly rise to a prohibitively high value, and the tube may be destroyed.

In the circuit of Fig. 17-3, the job of a current-limiting resistor is in part done by the upper section of the potentiometer R. But if the tube is not to be directly connected to the voltage source when the contact arm of the potentiometer is in its topmost position, it is important to add R_{lim}.

Because the gas tube and R_{lim} are series-connected, E_a is the sum of the voltage drops across the tube and the resistor:

$$E_a = V_a + V_R \qquad (17\text{-}1)$$

The voltage-current characteristic of a glow-discharge tube is shown in Fig. 17-4. The current is laid off as abscissa, and the voltage as ordinate for a better presentation of variations in voltage. Of course, the plot might be constructed the other way around, that is, with the current laid off as ordinate and the voltage as abscissa as is customary for vacuum tubes.

When the applied voltage is raised just above zero, a very weak current begins to flow. This is the dark discharge space, *I*. The dark-discharge current is very small, and it is plotted on a larger scale than the rest of the plot (in microamperes).

Point A is where a glow discharge occurs (this is the breakdown, starting or striking point). It corresponds to the striking potential V_{st}. A glow discharge is brought about abruptly. The minimum current at which a glow discharge is still possible is many times the dark-discharge current. The voltage drop across the tube decreases likewise abruptly by several volts or even more owing to a re-distribution of E_a between the d. c. resistance R_0 of the tube and R_{lim}.

During the dark discharge, R_0 is substantially greater than R_{lim} which is chosen so that a glow discharge can be produced. Under a dark discharge, practically all of E_a is impressed on the tube, and the voltage drop across R_{lim} is zero very nearly. When a glow discharge occurs, the current suddenly rises and produces a noticeable voltage drop across R_{lim}. Because of this, the voltage drop across the tube, V_a, is abruptly brought down. In other words, when a glow discharge occurs, R_0 is drastically reduced and becomes comparable with R_{lim}. The supply voltage E_a is re-distributed so that a marked proportion of it is dropped across R_{lim} and V_a is reduced to the same extent. Before a glow discharge, $V_a \approx E_a$. After a glow discharge has occurred, $V_a = E_a - i_a R_{\text{lim}}$. It is to be noted that E_a immediately before a glow discharge is practically the same as it is immediately after a glow discharge has occurred. The point is that if E_a is nearly equal to V_{st}, a very small increase in E_a is enough to cause a glow discharge.

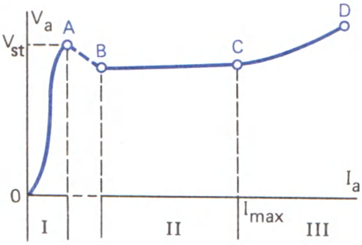


Fig. 17-4

Current-voltage characteristic of the dark discharge (region *I*) and the glow discharge (regions *II* and *III*)

Thus, the fact that a glow discharge has taken place is indicated by an abrupt rise in current and an abrupt fall in voltage as read from the instruments. There is also a glow of light at the cathode. On the plot, the occurrence of a glow discharge is represented by the region AB which cannot be plotted point-by-point; it can only be observed visually on an oscilloscope.

When plotting the voltage-current characteristic of a glow discharge tube, the voltage at point B, which is the maintaining potential for a glow discharge, is sometimes mistaken for V_{st}. Actually, however, the latter is the highest potential that can be observed when the supply voltage is raised before it falls abruptly. In contrast, the position of point B depends on the value of the current-limiting resistor. As the resistance of this resistor is reduced, the current rises, and point B shifts to the right.

After a glow discharge has been established, the rise in the applied voltage, E_a, is accompanied by an interesting event. The current rises, but the voltage drop across the tube increases only slightly until the current exceeds its maximum value, I_{max} (point C). This nearly constant voltage is known as the *normal cathode fall of potential* (region *II*). In this region (often called the *Crookes dark space*) the current flows through a part of the cathode's surface area, and there is a faint glow of light only in that part of the cathode. At a low current, only a small fraction of the cathode's total area is covered with the glow. When the current is increased, the area of glow on the cathode increases in proportion to the current, but the cathode current density remains unchanged. At I_{max}, all of the cathode's surface is covered with glow.

The normal cathode fall of potential is utilized in voltage regulator or voltage reference (VR) tubes. This mode of operation has a number of special features. Let the surface area of the cathode be substantially larger than that of the anode and let there be a suitable current-limiting resistor connected in the circuit (Fig. 17-5). Then, a relatively low current will be flowing around the circuit after a discharge occurs. This may be a glow discharge, provided the cathode current density is not too low. It is only then that a sufficient number of electrons can be knocked out of the cathode. The discharge does not cover all of the cathode's surface, and there is a current flow only through a part of it (shown shaded in the figure), so that the current density is high enough and a glow discharge can be maintained.

The voltage drop across the tube is

$$V_a = i_a R_0$$

where R_0 is the resistance presented by the ionized gas between the anode and the cathode's portion covered by the glow. This gas acts as a conductor which has the shape of a cone. If we raise the applied voltage, the current will increase, and the cathode's area covered by the glow will increase in proportion. As a result, the cross-sectional area of the gaseous 'conductor' will be greater, and R_0 will decrease to the same extent. Thus, R_0 decreases in the ratio as i_a increases and the product $i_a R_0$ remains constant (actually, it increases somewhat).

This condition will persist so long as the glow covers less than the total surface of the cathode. When the area of glow covers all of the cathode, a further rise in E_a will raise the current but the area of glow cannot be increased any longer. In this condition, an increase in the number of electrons knocked out of the cathode is only possible at the cost of an increase in the energy of the ions bombarding the cathode, and this calls for a higher applied voltage. The cathode current density is increased, but R_0 no longer decreases in proportion to the rise in current, and the product $i_a R_0$ (that is, the voltage drop across the tube) is increased, and there is then an *abnormal cathode fall of potential* (region *III* in the plot of Fig. 17-4).

Nevertheless, R_0 is somewhat decreased as the current rises because there are more ions and electrons per unit volume of the gas. However, this decrease is not so marked as in the case of the normal cathode fall of potential, and so V_a rises. The glow also gains in intensity and it extends through an ever greater volume of the plasma. The abnormal cathode fall of potential is utilized in gas-discharge lamps and indicator tubes.

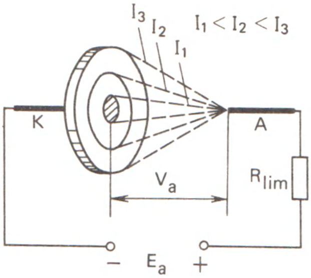


Fig. 17-5

Measuring the effective area of the cathode at the normal cathode fall of potential

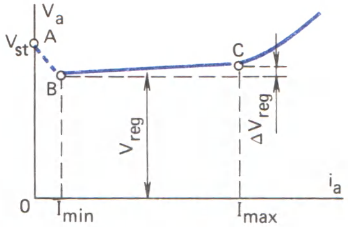


Fig. 17-6

Current-voltage characteristic of a VR tube

If we kept on rising the applied voltage, the current in and the voltage drop across the tube would build up until an arc would strike suddenly. However, an arc discharge would usually destroy a glow-discharge tube, so it must be avoided.

17-4 Voltage Regulator Tubes

Voltage regulator, or *voltage reference* (*VR*) *tubes* may be of the glow discharge type and of the corona discharge type. Use is most commonly made of glow-discharge VR tubes operating at the normal cathode fall of potential. Of late, they have been giving way to solid-state (crystal or semiconductor) breakdown diodes.

Because the dark discharge which precedes the glow discharge is not utilized, it is not shown on the *V-I* characteristic of VR tubes (Fig. 17-6). The point *A* at which the discharge is initiated is shown on the axis of ordinates. Practically, this is so because the milliammeter intended to measure the current of a glow discharge will fail to register the negligible current associated with a dark discharge.

The normal cathode fall of potential region used for voltage regulation (or stabilization) is bounded by the minimum current, I_{min}, and the maximum current, I_{max}. At a current smaller than I_{min}, the discharge may cease. The maximum current marks either the onset of the abnormal cathode fall of potential or the limit of power.

The abrupt rise in current secondary to a discharge may be different in magnitude, depending on the value of R_{lim}. If R_{lim} has a high value, a relative small current will flow. If, on the other hand, we take a high-value R_{lim}, a heavy current will flow, and point B will be shifted to point C. This is not suitable for voltage stabilization, because the region BC that can be utilized for this purpose is curtailed. If R_{lim} is too small, it may so happen that the tube will move into the abnormal cathode fall of potential region, and voltage stabilization will not be possible at all. Thus a well-chosen current-limiting resistor is needed for two reasons. Firstly, it prevents too high a rise in current. Secondly, it assures that the tube can operate as a voltage regulator.

The larger the cathode's surface area, the wider the stabilization region BC because I_{min} remains unchanged and I_{max} rises in proportion to the surface area of the cathode. Therefore, a typical VR tube has a cathode with a large surface area. The anode is made small, but still large enough to avoid overheating by I_{max}.

Most often, VR tubes have a cylindrical cathode made of nickel or steel. The anode is a piece of wire 1-1.5 mm in diameter (Fig. 17-7*a*). The envelope is filled with a mixture of inert gases (argon, helium, and neon), at a pressure of tens of mm Hg.

The principal parameters of VR tubes are the *normal operating* or *regulated output voltage*, V_{reg}, which corresponds to the middle point in the stabilization region (see Fig. 17-6), the striking voltage V_{st}, the minimum and maximum currents at the normal cathode fall of potential, I_{min} and I_{max}, the limit of variations in regulated voltage ΔV_{reg}, and the a.c. anode resistance r_{a}. If V_{reg} must be kept low, the inner side of the cathode is given a coat of an activator so as to facilitate the emission of electrons when the cathode is bombarded by ions. The desired value of V_{reg} is obtained by using a suitable mixture of inert gases. V_{st} usually exceeds V_{reg} by not more than 10-20 V. The value of V_{st} can be brought down by connecting a conductor (shown in Fig. 17-7*a*) to the inner side of the cathode. This conductor decreases the spacing between cathode and anode. Without it, the tube would operate within the up-sloping portion of the characteristic (see Fig. 17-2).

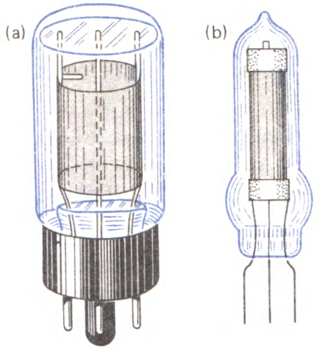


Fig. 17-7

VR tubes: (*a*) the glow-discharge type and (*b*) corona-discharge type

Within the voltage stabilization region V_{reg} may change by not more than ΔV_{reg} which usually is 1-2 V. It is not recommended to operate a VR tube at a current in excess of I_{max} because the voltage regulator action is then impaired and the electrodes are overheated. The a.c. anode resistance $r_{\text{a}} = \Delta v_{\text{a}}/\Delta i_{\text{a}}$ is markedly smaller than the tube's d.c. resistance, R_0. If voltage regulation were ideal (that is, $V_{\text{reg}} =$ const), r_{a} would be zero.

Soviet-made VR tubes are fabricated for a regulated output voltage of 75 V to several hundred volts. Their I_{min} is usually 3-5 mA, and their I_{max} is tens of milliamperes.

Voltage-regulator tubes of the corona discharge type are characterized by high voltages and small currents. They use cylindrical electrodes made of nickel (Fig. 17-7*b*). The envelope is filled with hydrogen, and the regulated output voltage depends on the gas pressure which usually is tens of mm Hg. Typical values of V_{reg} are several hundred volts. The operating currents lie in the range 3-100 μA. The a.c. anode resistance of these VR tubes is hundreds of kilohms. It takes from 15 to 30 s for a full corona discharge to develop.

The more recent makes of corona-discharge VR tubes are built into ceramic envelopes capable of operating at voltages as high as tens of kilovolts.

17-5 Circuits Using VR Tubes

In a practical circuit, a VR tube is connected in parallel with the load resistance R_L and in series with a current-limiting resistor R_{lim} (Fig. 17-8). The load may be the anode or screen-grid circuits of an amplifier which need a regulated (or stabilized) voltage. The supply voltage E must be in excess of V_{reg} and sufficient for a discharge to occur in the tube. Higher values of E call for higher values of R_{lim}. When this requirement is satisfied, voltage stabilization is maintained in the face of greater variations of E. When R_{lim} is too large, the efficiency of the circuit is impaired because the power dissipated in the tube and R_{lim} may exceed that drawn by the load. Therefore, VR tubes are only used for low-power equipments where a reduction in efficiency is not so important as it is in high-power loads.

Voltage regulator tubes are often operated when the load resistance is held at a constant value but the supply voltage is subject to variations. What happens in such cases may be broadly described as follows. When the supply voltage rises, the VR tube current increases, and nearly all of the supply voltage change is dropped across R_{lim}. The voltage across the tube and load remains nearly constant or rises only slightly, provided the tube current does not change beyond the limits for the normal cathode fall of potential. When E falls, the above events are reversed.

The value of R_{lim} can be found by Ohm's law. If E varies both ways from its average value, E_{av}, then

$$R_{lim} = (E_{av} - V_{reg})/(I_{av} + I_L) \quad (17\text{-}2)$$

where I_{av} = average current in the VR tube, equal to half the sum of I_{min} and I_{max}

I_L = load current defined as $I_L = V_{reg}/R_L$

The average supply voltage is found as the arithmetic mean of the minimum and maximum supply voltages:

$$E_{av} = (E_{min} + E_{max})/2 \quad (17\text{-}3)$$

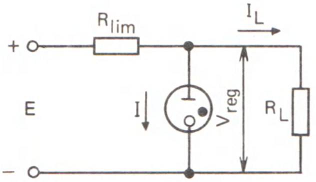


Fig. 17-8

Connection of a VR tube in a circuit

After R_{lim} has been calculated, it is important to check whether the VR tube will perform as required as the supply voltage varies from E_{min} to E_{max}. The check is run as follows.

When the tube current changes from I_{min} to I_{max}, the voltage drop across R_{lim} changes by

$$\Delta E = R_{lim}(I_{max} - I_{min})$$

The VR tube will perform well provided E changes by not more than ΔE. If $\Delta E < E_{max} - E_{min}$, the regulator action will affect only part of the range of changes in E, and this part will be the smaller, the lower the change ΔE.

Because I_{min} and I_{max} are constant for a given VR tube, ΔE is proportional to R_{lim}. However, the value of R_{lim} is increased in proportion to the difference between E and V_{reg} and in inverse proportion to I_L. Thus, a broader range of regulator action can be assured with a higher supply voltage and a lower load current. This however impairs the efficiency of the voltage regulator.

If the load current is large, R_{lim} is small, and the regulator action can be effected over a very narrow range of changes in E, which is obviously a disadvantage. Therefore, the use of VR tubes is warranted at I_L which does not markedly exceed I_{max}.

Higher voltages are usually regulated by connecting several VR tubes in series (as a rule, not more than two or three tubes are used). They may be rated for different regulated output voltages, but their I_{min} and I_{max} must be the same. A series connection of several VR tubes may be used as a voltage divider giving several values of regulated voltage. The loads may be connected to one or several VR tubes. For example, a voltage regulator made up of three 75-volt VR tubes may supply regulated voltages of 75, 150, and 225 V. Sometimes, the load voltage has to be different from the regulated voltages supplied by standard 75-, 105- or

150-volt VR tubes or their combinations. In such cases, one tube or several tubes are connected to give the nearest standard voltage, and the surplus is dropped across a swamping resistor, R_{sw}, connected in series with R_L (Fig. 17-9). For example, if we need a regulated voltage of 120 V at $I_L = 10$ mA, we may take a 150-volt VR tube and drop the surplus voltage (30 V) across a swamping resistor

$R_{sw} = 30/10 = 3$ kΩ

Parallel connection of VR tubes is never used because individual devices, even of the same type, usually differ in V_{st} and V_{reg}. When a parallel combination of VR tubes is energized, a discharge occurs only in the unit that has the lowest V_{st}. The voltage across that tube falls abruptly while no discharge happens in the remaining devices. Even if it occurred, some tubes would operate underloaded while the others would be overloaded. It might also so happen that some tube would be operating under an abnormal cathode fall of potential. It would then take no part in voltage stabilization – instead it would act as a useless load and would limit the range of voltage stabilization. Of course, it is possible to match the tubes for parameters, but this, too, is inconvenient and unreliable because the tube parameters vary with time, and do so at different rates in different tubes.

The performance of a voltage regulator can be stated in terms of the *stabilization factor* defined as the ratio of the fractional change in the regulated output voltage to the fractional change in the supply voltage

$$k_{reg} = (\Delta E/E)/(\Delta V_{reg}/V_{reg}) \qquad (17\text{-}4)$$

Typically, the stabilization factor k_{reg} ranges from 10 to 20. If, for example, $k_{reg} = 10$, $E = 200$ V, and $V_{reg} = 75$ V, then a change of $\Delta E = 40$ V in the supply voltage, that is, by 20%, will cause the regulated output voltage to change by a mere 1.5 V, or 2%.

The stabilization factor can be improved by connecting several VR tubes in cascade (Fig. 17-10). In this scheme, the voltage from the first tube VR_1 is applied via a current-limiting resistor R_{lim2} to the second tube VR_2 placed in shunt with the load. If the individual stabilization factors are k_{reg1} and k_{reg2}, the overall stabilization factor will be

$$k_{reg} = k_{reg1} k_{reg2} \qquad (17\text{-}5)$$

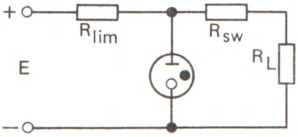


Fig. 17-9
Reduction in the regulated voltage by a series resistor

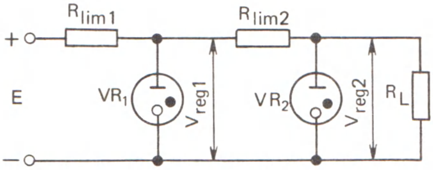


Fig. 17-10
Cascade connection of VR tubes

With two VR tubes in cascade, k_{reg} is anywhere from 100 to 400. A drawback of the cascade connection is a reduction in the efficiency because some power is inevitably lost in the two tubes and the two current-limiting resistors. More than two tubes are not usually used. The second tube must be designed for a lower voltage than the first, V_{reg1} may be assumed constant so that R_{lim2} may be calculated for the current through VR_2, which only slightly exceeds the minimum current.

Voltage regulator tubes may also be used in cases where the load resistance is varying while the supply voltage E remains constant. R_{lim} is then calculated as before. If I_L varies from its minimum value $I_{L\,lim}$ associated with $R_{L\,max}$ to $I_{L\,max}$ associated with $R_{L\,min}$, then

$$R_{lim} = (E - V_{reg})/(I_{av} + I_{L\,av}) \qquad (17\text{-}6)$$

where I_{av} = average current of the VR tube
$I_{L\,av}$ = average load current defined as half the sum of the minimum and maximum load currents

$$I_{L\,av} = (I_{L\,min} + I_{L\,max})/2 \qquad (17\text{-}7)$$

In the above condition, the current is divided between the VR tube and the load. For example, should the load current rise, the tube current will fall by about the same amount and V_{reg} will remain nearly constant. The total current will likewise remain nearly constant so that the voltage drop across R_{lim} will change only slightly. This is as should be expected because

$$V_{reg} + V_R = E = \text{const}$$

Of course, the stabilizing action will take place if the tube current does fall below $I_{\min}$ and does not rise above $I_{\max}$. The load current ought not to change by a greater amount than the maximum change in the tube current. Hence, the condition for stabilization may be written in the form of an inequality

$$I_{\mathrm{L\,max}} - I_{\mathrm{L\,min}} \leqslant I_{\max} - I_{\min} \tag{17-8}$$

Voltage regulator tubes present different resistances to d.c. and a.c. As the tube current varies, R_0 changes from units to tens of kilohms. For example, in the case of a VR tube for which $V_{\mathrm{reg}} = 150$ V, $I_{\max} = 30$ mA and $I_{\min} = 5$ mA, R_0 will vary from 5 to 30 kΩ. In contrast, r_{a} will be only a small fraction of R_0. Taking the same tube as an example, suppose that V_{reg} changes by 2.5 V as the tube current changes from 5 to 30 mA. Then

$$r_{\mathrm{a}} = \Delta V_{\mathrm{reg}}/\Delta I = 2.5/25 = 0.1\ \mathrm{k\Omega}$$

As seen by alternating current, a VR tube is equivalent to a high-value capacitor (at 50 Hz, a resistance of 100 Ω corresponds to a capacitance of 32 μF). Therefore, when used in rectifiers, VR tubes also double as ripple suppressors or, at least, as ripple reducers.

17-6 Glow-Discharge Thyratrons

Glow-discharge thyratrons, also known as *cold-cathode thyratrons*, are widely used tubes. They are employed in automatic control, relays, scalers, counters, pulsers, etc. The term 'thyratron' has its origin in the Greek *thyra* for door, thus implying that the tube can be opened and closed like a door with the aid of a control grid.

Thyratrons may have three and more electrodes. In a three-electrode glow-discharge thyratron, the anode and cathode are supplemented by a third electrode, called the *grid*. The grid of a thyratron effects a more limited control that it does in vacuum triodes. In a vacuum triode, we can vary the grid potential and thus control the anode current, that is, cause it to vary from zero to a maximum value. In a thyratron, the grid has a one-way control of conduction and only serves to fire the tube at the instant when it acquires a critical voltage. Following that, the grid effects no control over conduction – the discharge in a thyratron can be discontinued only by reducing the anode voltage to a value at which no discharge can be sustained, or by opening the anode circuit.

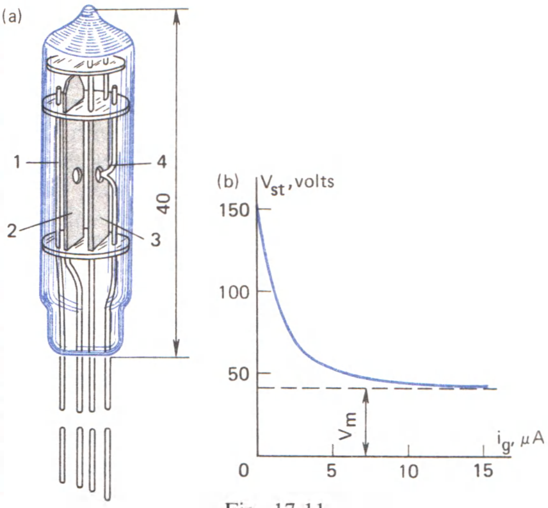


Fig. 17-11

Structure and firing characteristic of a glow-discharge thyratron: (*1*) anode; (*2*) 2nd grid; (*3*) 1st grid; (*4*) cathode

Figure 17-11*a* shows the structure of one type of glow-discharge thyratron. The separation between the electrodes and the gas pressure are chosen such that a self-sustained dark discharge can occur between the grid and cathode at a lower voltage than between the anode and cathode. It may then be followed by a glow discharge to the anode, provided the anode voltage is high enough. The grid current is units or tens of microamperes, and the anode current may be thousands of times greater (units or tens of milliamperes). The *starting* or *firing potential*, V_{st}, decreases with increasing grid current i_{g}. This is because more ions and electrons are produced in the grid-cathode space at a high grid voltage, and it is easier for a discharge to take place in the anode circuit.

A plot of V_{st} as a function of i_{g} (Fig. 17-11*b*) is usually called the *starting* or *firing characteristic* of a thyratron. At zero grid current, the starting or firing potential is a maximum. As i_{g} rises, V_{st} falls, abruptly at first and more gradually afterwards. However, V_{st} cannot be lower than the *maintaining potential* V_{m} required to sustain a glow discharge between the anode and cathode. The shape of the starting curve depends on the gas used, its pressure, electrode shape, and electrode surface texture.

The fact that the grid loses control after the tube has fired can be explained by the existence of a plasma made up of a large number of electrons and ions. The positively charged grid attracts electrons from the plasma, and they form a *negative* (or *electron*) *sheath* at the surface of the grid, neutralizing the action of the positive grid (Fig. 17-12*a*). If we raise or lower the positive potential at the grid, it will attract a greater or a smaller number of electrons, and its effect will again be neutralized by the electron sheath. A negative voltage across the grid will attract positive ions from the plasma, which will form a *positive-ion sheath* around the grid, neutralizing the effect of the negative grid (Fig. 17-12*b*).

The electron (or ion) sheath is in a state of dynamic equilibrium. For example, on coming in contact with the negative grid, the ions capture electrons and turn into neutral atoms, but the grid attracts more ions from the plasma instead of them. If we raise the negative potential at the grid, it will attract more ions. The charge on the ion sheath will rise, and the ion sheath will again neutralize the effect of the negative grid. It may be said that the field set up by the charge on the grid is concentrated between the grid and its ion (or electron) sheath as if they were the plates of a capacitor. This field cannot break through the sheath, and so it cannot affect the anode current.

Figure 17-13 shows how a glow-discharge thyratron can be connected in a circuit for use as a relay. The anode supply voltage E_a must be lower than $V_{st\ max}$, and E_g must be lower than that required to start a discharge between the grid and cathode. The resistor R_g limits the grid current and thus raises the input resistance of the circuit to the source of pulses that turn on the thyratron. When a positive voltage pulse strong enough to fire the tube is applied to the grid, a discharge takes place between the grid and cathode. If the resultant grid current is large enough, the discharge is transferred to the anode. In consequence, a voltage or current pulse supplied by a low-power generator to the grid circuit can give rise to a substantial current in the load, R_L, placed in the anode circuit.

Some glow-discharge (cold-cathode) thyratrons have two grids. The one farthest from the cathode is the control grid, and the one nearest to the cathode is the screen grid. The screen grid is held at a constant positive potential, and its circuit always draws a very small current (units or tens of microamperes) used for a starting discharge. The control grid is less positive than the screen grid, therefore, the retarding field existing between the grids does not allow electrons to reach the anode. When an additional positive voltage pulse is applied to the control grid, the thyratron is said to fire – electrons pass through the control grid, and a glow discharge is initiated at the anode.

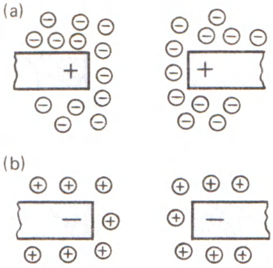


Fig. 17-12

Electron and ion sheaths at the grid

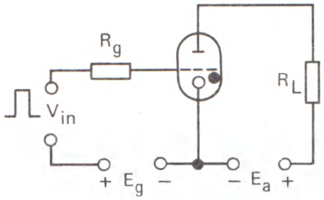


Fig. 17-13

Connection of a glow-discharge thyratron as a relay

Soviet-made glow-discharge thyratrons are usually miniaturized and use neon or argon or their mixture as the filling gas. They are able to operate at an ambient temperature of −60° to +100°C. Their service life is several thousand hours. The operating grid and anode voltages range from tens of volts to 100-300 volts. The time required for the grid to regain its control action after the anode current ceases depends on the de-ionization time and usually is tens or hundreds of microseconds.

As an example, Fig. 17-14*a* shows the circuit of a very simple thyratron sawtooth voltage generator. The anode supply, E_a, charges a capacitor C via a resistor R. The capacitor is placed in shunt with a thyratron T. In the course of charging, the voltage across the capacitor rises until it comes up to V_{st}, the starting or striking voltage for the thyratron. The thyratron fires and begins to conduct current. Its resistance falls to a relatively low value, and the capacitor rapidly discharges through the thyra-

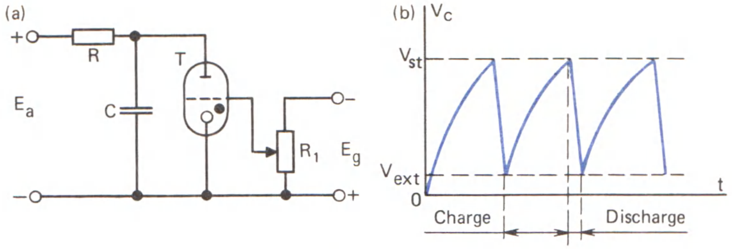


Fig. 17-14

Sawtooth voltage generator built around a thyratron

tron. The voltage across the tube falls to its *extinction value*, V_{ext}, and the discharge ceases. Just as this happens, the capacitor begins to charge again slowly via the resistor whose resistance is substantially greater than that of the conducting thyratron, and all the events described above are repeated again.

The waveform of the sawtooth voltage generated at the anode of the thyratron and across the capacitor is shown in Fig. 17-14*b*. Because the extinction voltage, V_{ext}, of a thyratron is small and the striking voltage, V_{st}, is several hundred volts, this type of oscillator can generate a sawtooth voltage of a high peak value. An increase in the values of R and C slows down the charging process and the frequency of the output voltage is reduced. On the contrary, an increase in the positive potential at the grid brings down V_{st}, thus leading to a lower amplitude and a higher frequency of the sawtooth voltage.

17-7 Character Display and Numerical Readout Devices

Character display and numerical readout devices are widely used in electronics. Some of them fall in the class of glow-discharge tubes but there are also vacuum devices doing similar jobs. Of late, there has been a growing trend to use semiconductor (or *solid-state*) character display and numerical readout devices; they will be discussed in more detail in Chap. 23.

Neon bulbs. A neon bulb is essentially a glass envelope filled with neon gas and containing two (or more) insulated electrodes. These devices are usually employed as voltage indicators but they can do some other jobs as well. Two-electrode neon bulbs usually operate at the abnormal cathode fall of potential and always in conjunction with a current-limiting resistor, R_{lim}.

The current-voltage characteristic of a neon bulb is shown in Fig. 17-15. When a discharge occurs (point A), the current and voltage change abruptly, and there is a glow of light. The further rise in voltage leads to a rise in current. This in turn brings about an increase in the cathode current density and a brighter glow. Characteristically, a decrease in voltage causes the curve to run lower than it does when the voltage is raised. Changes in the current through the bulb lag behind the changes in the voltage across its electrodes, and the discharge goes out at a lower voltage than it strikes, that is, $V_{\text{ext}} < V_{\text{st}}$. Just as the discharge ceases, the current suddenly falls to zero, and the voltage as suddenly rises because the voltage drop across R_{lim} abruptly falls to zero and the applied voltage E is divided in a different way. Experimentally, V_{ext} is defined as the lowest voltage existing in the presence of current and glow in the bulb (prior to the extinction of the discharge).

The difference between V_{ext} and V_{st} is typical of all gas-discharge devices, notably VR tubes. The extinction voltage of neon bulbs is by several units or tens of volts lower than the striking voltage. This is because the gas is not ionized before a discharge occurs. In contrast, before the extinction the gas is ionized, and a discharge can exist at a lower voltage.

Neon bulbs can be used as a.c. and d.c. voltage indicators. With a.c., a discharge takes place just as the instantaneous voltage becomes equal to V_{st}.

Commercially available neon bulbs can be made for V_{st} ranging from 50 to 200 V or even higher. The operating current at normal glow is from a few tenths of a milliampere to tens of milliamperes.

Of special interest is a three-electrode indicator tube which has an anode, an indicator cathode, and an auxiliary cathode enclosed in the anode. The observer can see the glow of gas only at the indicator cathode through the dome of the envelope. The indicator cathode, *IC*, is connected to the "−" terminal of a power supply via a resistor *R*, while the auxiliary cathode *AC* is connected to the supply directly (Fig. 17-16). When only the anode voltage is applied to the tube, the auxiliary cathode is operating. Because it is in the shadow of the anode, the observer cannot see the glow of gas. When an additional control voltage is applied to the resistor in the indicator cathode circuit in a polarity such that it is added to the anode voltage, the voltage between the anode and the indicator cathode rises, the discharge is transferred to the indicator cathode, and the observer is able to see the glow of light. When the additional voltage is removed from the resistor, the discharge will again exist only between the anode and the auxiliary cathode, and there will be no glow of light at the indicator cathode.

Glow-discharge character-display tubes. The structure and graphical symbol of these widely used devices are shown in Fig. 17-17. Such a tube is essentially a neon-filled envelope enclosing a common anode and stacked metallic cathodes in the form of the desired characters. For simplicity, Fig. 17-17*a* shows only the first two cathodes in the form of a 1 and a 2. A numerical-readout tube will have the numerals 0 through 9. The anode is usually a fine mesh of wire. When a voltage is applied between the anode and one of the cathodes, a gas discharge takes place and a glow of light can be seen at that cathode, so that the observer can see the respective character. The luminous bar is about 1 or 2 mm wide. Another widely used display format consists of *segments* arranged as in Fig. 17-17*b*. Connection of these segments in a particular combination produces the desired luminous numeral, letter, or some other symbol.

Vacuum hot-cathode character-display tubes. The display format in this case consists of incandescent filaments that are combined to produce a numeral or a letter (Fig. 17-18). These filaments made of tungsten are set up on a heat-resistant insulating support inside an evacuated envelope. At one end the filaments are tied to a common lead. When the other end of a particular filament is connected to the heater supply, the luminous segments form a numeral or a letter. The glow colour is yellow which corresponds to an operating temperature of 1200°C. The service life of vacuum hot-cathode character-display tubes is tens of thousands of hours.

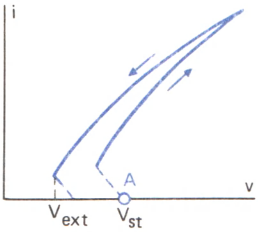


Fig. 17-15

Current-voltage characteristic and graphical symbol of the neon bulb

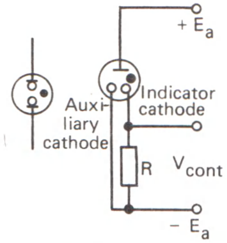


Fig. 17-16

Connection of a controlled indicator tube

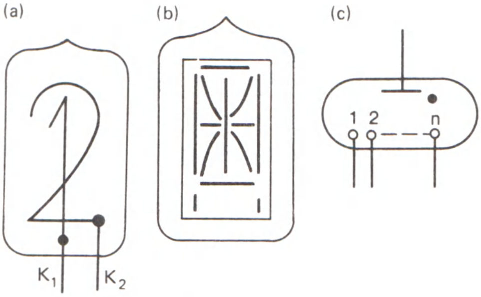


Fig. 17-17

Glow-discharge character readout tube

Vacuum fluorescent indicators. These are multi-anode triodes each of which has a directly heated oxide cathode, a grid, and phosphor-coated segments as anodes. The anodes may be set up in any one of several patterns to form the desired character (Fig. 17-19). The emission colour is mostly green.

Electroluminescent devices. For their operation these devices depend on *electroluminescence* – a luminescence which results from a high-frequency discharge through a gas or from application of an alternating current to a layer of phosphor. A solid-state electroluminescent device looks like a flat (parallel-plane) capacitor (Fig. 17-20). There is a metal electrode *1* to which is deposited a layer of dielectric *2*, a mixture of organic resin and powdered phosphor. The phosphor is basically zinc sulphide or zinc selenide. Depending on the activator added, the phosphor will give a different emission colour: green, blue, yellow, red, or white. The phosphor is topped by a current-conducting film *3*. Protection against external damage is provided by a glass cover plate *4*. When an alternating current is applied to the electrodes *1* and *3*, an emission of light takes place under the action of the electric field.

The transparent electrode *3* is usually prepared from tin oxide as a single piece. The other electrode, *1*, may be shaped into numerals, letters, or segments which can then be combined to present symbols or shapes. The electrode *1* may be the *raster type* (consisting of a number of stripes) or the *matrix type* (consisting of a great number of dots). Electroluminescent devices come in a wide range of types and sizes. They can display a luminous pattern against a dark background or a dark pattern against a luminous background. Also, they may be single- or multi-coloured.

Most commonly, use is made of segment-type alphanumerical readout devices. Digital-indicator devices have from seven to nine segments, while devices with as many as 19 segments can display any numerals and letters of one or two alphabets (say, Russian and English). As a rule, an electroluminescent readout device is built into a plastic case. They are powered by a sinusoidal current at 220 V and 400-1200 Hz. The characters displayed by electroluminescent devices can measure from units to tens of millimetres. Accordingly, they will draw a current of a few tenths of a milliampere to tens of milliamperes. The service life of an electroluminescent readout device is several thousand hours. They can operate at an ambient temperature of −40 to +50°C. Among the most important advantages of these devices are low power drain, a relatively high brightness of the display, flat construction, mechanical robustness, and a long service life. A major disadvantage, common to many character-display devices, is the need for a sophisticated control system.

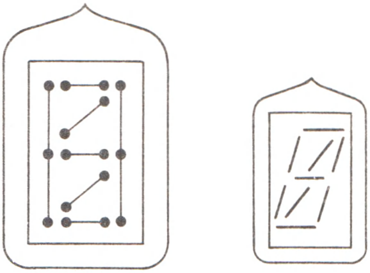


Fig. 17-18

Vacuum hot-cathode character-display tube

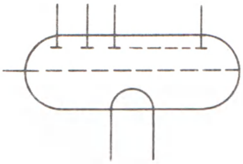


Fig. 17-19

Vacuum fluorescent character-display tube

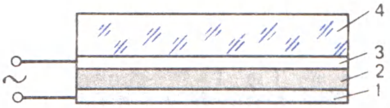


Fig. 17-20

Electroluminescent readout device

Liquid-crystal displays (LCD). What we know today as *liquid crystals* were discovered at the end of the past century. They are organic liquids consisting of long-chain molecules that line up under the influence of an applied electric field to give a quasi-crystalline structure to the liquid. At this writing, quite a number of liquid-crystal materials have been found and investigated. Liquid crystals are transparent to visible light, but an electric field with an intensity of 2 to 5 kV cm^{-1} disturbs their regular structure, the molecules arrange themselves at random, and the liquid becomes opaque. This is the basis of their operation as passive display devices.

LCDs may be of many designs and operate either by transmitting or reflecting light from a suitable source (natural or artificial). Most commonly, LCDs are of the reflecting type (Fig. 17-21) and are widely employed in digital watches, pocket calculators, and other products. Referring to the figure, there are two glass plates, *1* and *3*, cemented together by a polymer resin

2 and holding between them a layer of liquid crystal *4* from 10 to 20 μm thick. Plate *3* is given a coat of a solid conducting layer (an electrode), *5*, worked to a mirror finish. The other plate, *1*, gives support to transparent electrodes *A*, *B*, *C*, etc., from which leads are made (not shown in the figure). These electrodes are shaped as various numerals, letters, or segments which can be combined into a particular character. So long as no voltage is applied to the character electrodes, the LCD remains transparent, external light is free to pass through and to be reflected from electrode *5*, and to emerge without displaying any symbols. When a voltage is applied to, say, electrode *A*, the liquid crystal under this electrode turns opaque, light cannot any longer pass through that part of the liquid (*6*), and a dark character can be seen against an illuminated background.

LCDs are very economical of power. It takes not more than 1 μA to display one character. Their service life is tens of thousands of hours. Unfortunately, LCDs are slow-acting devices—it takes 100-200 ms for a character to appear or disappear, that is, for the molecules of a liquid crystal to change from an ordered to a disordered arrangement or back. Also, control of LCDs is effected by sophisticated circuits usually based on IC chips.

The range of LCDs available commercially includes far more sophisticated designs than we have described.

17-8 Miscellaneous Gas-Discharge Devices

Quite a number of gas tubes have fallen out of use at this writing. Yet, they are worth mentioning.

One example is the *direct-counting decade tube* of the Dekatron type developed by Ericsson Telephone, Ltd. This type of tube has a great number of cathodes arranged in a circle. Incoming pulses transfer the discharge from one cathode to the next higher one. The number of pulses counted by the tube is indicated by the glow of the respective cathode. With several decade tubes connected in cascade, it is possible to count units, tens, hundreds, thousands, etc. of pulses. In such a case, when the tenth cathode in the units tube comes aglow, the discharge is transferred to the first cathode of the tube in the tens position, etc. At this writing, decade tubes have given way to solid-state digital readout devices.

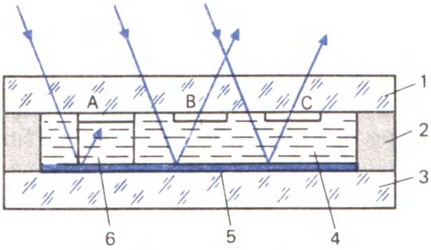


Fig. 17-21

Structure and operation of a liquid-crystal device

Arc-discharge devices also include *high-power hot-cathode rectifier diodes* whose envelopes are filled with mercury vapour or inert gases. They are designed for heavy-duty applications (that is, high voltages and heavy currents), with the voltage drop across a tube being a mere 10-30 volts. Falling also in this class are *mercury-arc rectifier tubes* and *single-anode pool tubes* with means for maintaining a continuous cathode spot (known as *excitrons*), owing to electrostatic emission. Better performance is shown by *ignitrons*, a type of mercury-pool rectifier tubes which has only one anode. The arc is started for each cycle of operation by an *ignitor* which dips into the mercury pool. The mercury pool serves as the cathode of the tube.

At one time, automatic control and some other applications relied heavily on *arc-discharge thyratrons*, that is, gas-filled hot-cathode triodes. As in glow-discharge (cold-cathode) thyratrons, the grid loses control over the anode current as soon as the tube fires, so it can only turn the tube on and hold it in the OFF state. Some hot-cathode thyratrons have a second (or screen) grid. By varying the potential at that grid, we can adjust the starting (or striking) voltage. Arc-discharge thyratrons are used in controlled rectifiers where the rectified voltage is regulated by varying the voltage applied to the tube's grids. These grids draw very little power for control purposes, so the efficiency is very high. Strong but short voltage pulses can be produced by *pulsed arc-discharge thyratrons.*

A later modification of the arc-discharge thyratron is the *Tacitron* in which the grid is designed not only to initiate, but also to extinguish the discharge. Still another arc-discharge thyratron is the *Arctron* in which the cathode is heated by ion bombardment rather than by the filament current.

All of the devices listed above suffer from a noticeable time lag because the recombination that takes place after the tube has been turned off needs an appreciable time to reach completion. Because of this, such devices cannot be operated at high frequencies. Tubes filled with inert gases can operate at tens of kilohertz, and those filled with mercury vapour at still lower frequencies.

Chapter Eighteen
Tube Noise

18-1 Sources of Tube Noise

When the tubes of a receiver are operating at a high gain, a characteristic hissing, crackling or similar sound can be heard on the headphones or the speaker plugged into the receiver's output, even if no signals are applied to its input.

Indeed, this form of noise can be heard on any radio receiver if we disconnect its antenna and short-circuit its input terminals so that no external signals could enter it. This inherent receiver noise grows in strength as we use a higher gain. The primary cause of this noise lies in fluctuations. After amplification, these fluctuations manifest themselves as audible noise when the incoming signals are received by ear.

Inherent tube noise limits the sensitivity of radio receivers and other electronic circuits intended to detect, amplify and measure or otherwise display weak electrical signals. When incoming signals are weaker than the inherent noise, their reception by traditional methods is practically unfeasible.

The basic cause of inherent tube noise is likewise various fluctuations.

1. *Fluctuations in the electron emission from the cathode* may be caused by several factors. The number of electrons emitted by the cathode during identical time intervals does not remain precisely constant. Therefore, the emission current is continuously oscillating in a haphazard fashion even if the emitter surface remains unchanged. More commonly, this is known as *shot noise* (also called *Schottky noise* after the German scientist who was the first to observe and describe the phenomenon).

The emissivity of microscopic areas on the surface of a cathode is likewise subject to continuous, rapid and random changes. The result has come to be known as the *surface fluctuation effect*.

The above two effects are observed with any form of emission and with any type of cathode, but to a different degree. They are most felt in the case of thermionic emission and activated cathodes. The surface fluctuation effect may be especially strong with oxide cathodes.

2. *Fluctuations in the secondary electron emission from tube's electrodes* held at a positive potential, insulators, and glass envelope also contribute to inherent tube noise.

3. *Fluctuations in ion currents* are observed in soft or gassy (that is, imperfectly evacuated) vacuum tubes. As the quality of the vacuum deteriorates, increasingly more ions are produced, and the effect due to this type of fluctuations grows stronger.

4. *Fluctuations in current division* occur always when a tube has two or more electrodes held at positive potential. Owing to thermal chaotic motion, the number of electrons hitting these electrodes is varying continuously and in a random manner giving rise to *partition noise.*

18-2 The Noise Performance of Vacuum Tubes

When a vacuum diode is in the temperature-limited region (at saturation), the rms value of its noise current may be found by the following equation:

$$I_n^2 = 2eI_{sat}B \tag{18-1}$$

where e = charge on an electron
I_{sat} = saturation current
B = bandwidth of the circuit used to observe the noise current

For example, if $I_{\text{sat}} = 50$ mA and $B = 1$ kHz, then

$$I_n = (2 \times 1.6 \times 10^{-19} \times 50 \times 10^{-3} \times 10^3)^{1/2}$$
$$= 4 \times 10^{-9}\ \text{A} = 4 \times 10^{-3}\ \mu\text{A}$$

In the space-charge-limited region, the noise current decreases. This happens for the following reason. Let the emission increase somewhat due to fluctuations. This means that a greater number of electrons are emitted from the cathode per unit time. The anode current should, it would seem, rise in magnitude. However, the space charge builds up as well, giving rise to a higher potential barrier at the cathode and thus leading to a decrease in anode current. Thus, we have two mutually opposing effects, and so the anode current fluctuates less in the space-charge-limited region than in the temperature-limited region (at saturation).

Since the noise current of a diode in the temperature-limited region can readily be found by the above equation, it is widely practised to use specially designed noise diodes as noise generators when testing radio equipment and other electronic circuits.

Different tubes are compared for noise performance by invoking the equivalent noise voltage $V_{\text{n eq}}$ and the noise resistance $R_{\text{n eq}}$ of tubes. These two quantities are derived from the following considerations.

It is assumed that the tube itself is ideal, that is, free from noise, but it generates noise due to some noise voltage applied to its grid. This is a voltage which will produce the same noise as the tube at room temperature and within a bandwidth of 1 kHz, and is called the *equivalent noise voltage* of a given tube. In other words, we may assume that the tube is an ideal (noise-free) one, and its grid lead contains a generator of $V_{\text{n eq}}$ (Fig. 18-1). For most tubes, $V_{\text{n eq}}$ is a fraction of a microvolt. Over a bandwidth of B kilohertz, the noise voltage is $B^{1/2}$ times the value of $V_{\text{n eq}}$.

Noise voltage is generated across any resistor. By Nyquist's equation (see Chap. 6), this voltage at room temperature is

$$V_n \approx (RB)^{1/2}/8 \tag{18-2}$$

where V_n is in microvolts, R in kilohms, and B in kilohertz.

It may be assumed that the noise voltage of a tube is generated by a resistor of resistance $R_{\text{n eq}}$ placed in the grid lead of the tube (Fig. 18-2). Because $V_{\text{n eq}}$ is defined at $B = 1$ kHz, the relationship between $V_{\text{n eq}}$ in microvolts and $R_{\text{n eq}}$ in kilohms may, in accord with Eq. (18-2), be written as

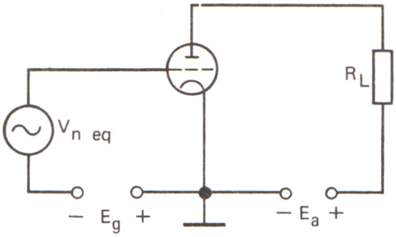


Fig. 18-1
Equivalent noise voltage of a tube

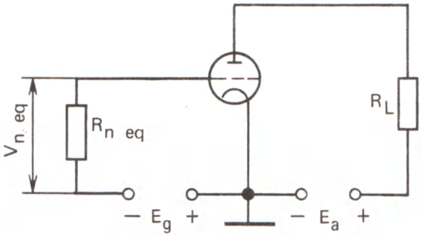


Fig. 18-2
Equivalent noise resistance of a tube

$$V_{\text{n eq}} \approx (R_{\text{n eq}})^{1/2}/8 \tag{18-3}$$

or

$$R_{\text{n eq}} \approx 64\, V^2_{\text{n eq}} \tag{18-4}$$

It is convenient to characterize the noise performance of tubes in terms of the *equivalent noise resistance* because one can readily calculate the total noise produced by the tube and the associated circuit components, such as resistors, in the grid lead.

The value of $R_{\text{n eq}}$ in kilohms for various tubes can be found by the following equation.

For a vacuum triode:

$$R_{\text{n eq}} \approx 2.5/g_m \tag{18-5}$$

For a pentode or tetrode:

$$R_{\text{n eq}} \approx 2.5/g_m + 20 I_{\text{a}} I_{\text{g2}}/g_m^2 (I_{\text{a}} + I_{\text{g2}}) \tag{18-6}$$

In the above equations, the currents are in milliamperes and the transconductance is in milliamperes per volt.

It is seen from the above equations that a reduction in $R_{\text{n eq}}$ can be achieved by increasing the transconductance of the tube. For triodes

$R_{\text{n eq}}$ is hundreds or even thousands of ohms; for pentodes and tetrodes it is higher (tens of kilohms), as more noise comes as partition noise. This resistance has still higher values for multigrid frequency-changing tubes. Generally, the noise level always rises with increasing number of electrodes in a tube. If a receiver or an amplifier is to have the lowest attainable noise level, the first stage should use a tube with the lowest possible value of $R_{\text{n eq}}$ because the noise originating in the first tube will be amplified by all the succeeding stages.

Sometimes the noise performance of tubes is stated in terms of their *noise figure* which is defined for tubes in the same way as for transistors (see Chap. 6). The magnitude of noise strongly depends on the operating conditions of a tube. A decrease in filament voltage leads to a stronger noise because this brings about a reduction in the space charge which somewhat smoothens fluctuations in anode current. An increase in the negative grid bias voltage is accompanied by a rise in tube noise because the transconductance of the tube is reduced. The same happens when the potential at the screen grid is reduced by a large amount. On the other hand, when V_{g2} is raised appreciably, more noise is produced owing to a change in current division (a greater contribution comes then from partition noise). There is an optimal value of V_{g2} at which noise is a minimum. Pentodes are less noisy in the intercept mode, while in the fall-back mode the transconductance is reduced and a greater contribution comes from partition noise. In operation at lower frequencies, the surface fluctuation effect is fairly strong. Thus, noise can be minimized not only through the choice of a tube, but also by adjusting its operating conditions (voltages and currents), as the case may be.

Noise due to fluctuations may be accompanied by noise due to other sources inside the tube itself. Among them are the *hum* that arises when the filament is energized with an alternating current, variations in current due to mechanical vibration of the tube's electrodes, current leakage through the insulation having a varying resistance, etc.

Chapter Nineteen
Operation of Vacuum Tubes at Microwave Frequencies

19-1 The Effects of Interelectrode Capacitances and Lead Inductances

There is always a capacitance between any two electrodes of a tube, and any tube lead has an inductance. As an example, Fig. 19-1 shows a triode along with its capacitances and inductances and an applicable equivalent circuit. These capacitances and inductances affect the parameters of the resonant circuits connected to the tube. As a result, the natural frequency of these circuits is reduced and they cannot be tuned to a frequency above a certain limit.

For each tube there is an *upper frequency limit*, f_h, which is set when the resonant circuit is produced by tying together the electrode leads of the tube. For example, if we tie together the anode and grid of a triode as shown by the dashed line in Fig. 19-1, a resonant circuit will be formed whose capacitance and inductance are given by

$$C = C_{ag} + C_{ak} C_{gk}/(C_{ak} + C_{gk}) \qquad (19\text{-}1)$$

$$L = L_a + L_g + L_{sc} \qquad (19\text{-}2)$$

where L_{sc} is the inductance of the short-circuiting wire.

A tube with an external resonant circuit can only operate at frequencies lying below f_h. As an example, let us choose a tube for which $C = 10$ pF and $L = 0.016$ μH. Then

$$\begin{aligned} f_h &= 1/[2\pi (LC)^{1/2}] \\ &= 1/[2\pi (0.016 \times 10^{-6} \times 10 \times 10^{-12})^{1/2}] \\ &\approx 400 \times 10^6 \text{ Hz} = 400 \text{ MHz} \end{aligned}$$

which corresponds to a wavelength of 75 cm.

Obviously, the above tube is unsuitable for use in the UHF band, because an external resonant circuit would require a resonant frequency well below 400 MHz.

The tube inductances and capacitances, when lumped together with the tube circuits, produce undesirable positive or negative feedback and phase shifts that would impair the performance of the entire circuit in many cases, if measures were not taken to the contrary. This is especially true of the cathode lead inductance L_k. It enters both the anode and grid circuits and gives rise to feedback owing to which the operating conditions are changed and there is a reduction in the input impedance of the tube, that is, the impedance between grid and cathode which loads the signal source. The interelectrode capacitances, too, tend to bring down the input impedance of the tube. Furthermore, since their reactance at microwave frequencies is very small, these interelectrode capacitances might give rise in high-power tubes to substantial capacitive currents which might heat the electrode leads and dissipate more power. For example, a grid-to-cathode capacitance of 4 pF will have a reactance of 40 Ω at 1 GHz ($\lambda = 30$ cm). If we apply an alternating voltage of 40 V, the capacitive current will be as heavy as 1 A.

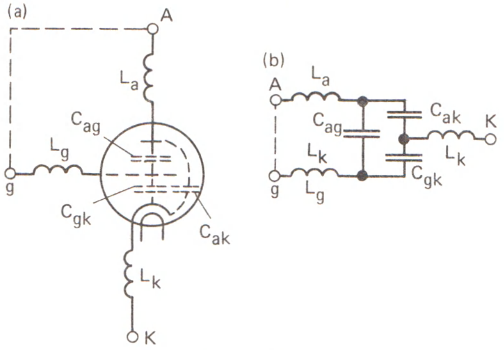


Fig. 19-1

Interelectrode capacitances and lead inductances of a triode

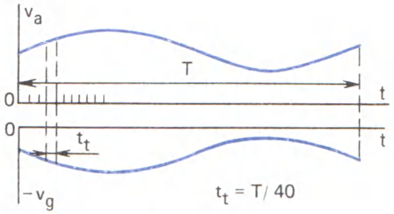


Fig. 19-2

Transit time of electrons as compared with the period of oscillations

19-2 Electron-Inertia Effects

Because electrons have a finite mass, they cannot cover the spacing between the electrodes instantaneously. At microwave frequencies, the transit time for electrons, although very short (10^{-8}-10^{-10} s), is comparable with the period of oscillation. The tube ceases to be a device free from a time lag. At microwave frequencies, the operation of triodes and multigrid tubes is said to be affected by *electron-inertia effects*. Electron inertia is responsible for detrimental phase shifts, distorts the waveform of anode-current pulses, and gives rise to appreciable grid currents. As a net result, the input impedance of the tube is drastically reduced, more power is dissipated in the tube, and the tube delivers less useful power.

Electron inertia does not affect the operation of tubes at frequencies lying in the VHF and higher frequency bands. The point is that if the period of oscillation, T, is many times the transit time of electrons in a tube, t_t, the alternating voltages maintained at the tube electrodes will have no time to change by a large amount at the end of electron transit. This is illustrated in Fig. 19-2 which shows how the voltages at the grid and anode of an amplifying tube vary when the period of oscillation is 40 times the transit time of electrons. For example, if $t_t = 10^{-9}$ s, then $T = 40 \times 10^{-9}$ s, which corresponds to

$$f = 1/(40 \times 10^{-9}) = 25 \times 10^6 \text{ Hz} = 25 \text{ MHz}$$

That is, the wavelength will be $\lambda = 12$ m.

In the above conditions we may take it that the voltages at the electrodes remain unchanged as electrons travel from cathode to anode. This means that the motion of electrons obeys the usual laws without any new events, so the anode current varies in step with the grid voltage, and the alternating component of anode current is in phase with the alternating component of grid voltage. The situation is markedly different when the transit time of electrons in a tube is of the same order of magnitude as the period of oscillation.

When the voltages at the electrodes of a tube remain unchanged, the tube is said to be in the *static mode of operation*. If, on the other hand, the voltage of at least one electrode varies at not too high a frequency (so that the processes in the tube may be treated on the basis of the laws established for the static mode), we will have a *quasi-static mode of operation*. Finally, if the voltage of at least one electrode varies so fast that the laws established for the static condition are no longer applicable, we have the *dynamic mode of operation*. At microwave frequencies, tubes operate in the dynamic mode. The laws that hold for the static condition cannot be applied to the dynamic condition because of electron inertia.

The transit time of electrons is often replaced with the *transit angle* α_t which is connected to the transit time by a relation of the form

$$\alpha_t = \omega t_t \qquad (19\text{-}3)$$

where ω is the angular frequency of the alternating voltage at the tube's electrodes.

Obviously, α_t is the change in the phase angle of the alternating voltage during t_t. If, for example, $t_t = T/4$, then $\alpha_t = 90°$. With transit angles of less than 20°, electron inertia may usually be neglected, and the tube may be regarded as operating in a quasi-static mode.

Let us see what happens in a vacuum triode at microwave frequencies, recalling that the greater proportion of the transit time is expended between cathode and grid because the accelerating potential difference there is small. Assume, as an example, that the transit time within that space is equal to a half-period and that the Q-point is positioned at the very start of the anode-grid characteristic of the tube. At lower frequencies, the tube would be at cutoff, that is, anode current pulses would be passed during the positive half-cycles of alternating grid voltage, and they would be blocked (the tube would be cut off) during the negative half-cycles.

A different situation emerges when $t_t = T/2$. The electrons setting out on their journey from the cathode at the start of a positive half-cycle of grid voltage pass through the grid at the end of the same half-cycle. The succeeding electrons setting out at a later time fail to reach the grid during a positive half-cycle. They are still on their way when the alternating grid voltage changes sign and the field between the grid and cathode becomes retarding. Many electrons are slowed down, brought to a stop, and driven back to the cathode before they have reached the grid. This is especially true of the electrons that set out on their travel from the cathode at the end of a positive half-cycle, because they find themselves in a retarding field almost at once. The return of some electrons to the cathode reduces the height of anode-current pulses. The tube delivers less useful power, and the returning electrons bombard the cathode. Because of this, the cathode is additionally heated, and this additional heat has to be supplied by the signal source. On the contrary, the electrons that have passed through the grid can continue their travel to the anode while the grid becomes negative. This means a greater potential difference between anode and grid, and so the electrons bombard the anode at an increased energy. The power expended in this bombardment is likewise drawn from the signal source.

If we consider the operation of a tube under different sets of conditions, we will arrive at the same conclusion: Electron inertia tends to reduce the alternating component of anode current, to increase the power dissipated at the anode, and to heat the cathode due to the action of returning electrons. These events happen not only when $t_t = T/2$, but whenever the transit time is comparable with the period of oscillation.

19-3 Electrostatically Induced Currents in the Circuits of Vacuum Tubes

For proper insight into the operation of vacuum tubes at microwave frequencies, it is important to learn what currents are electrostatically induced in the various circuits of these tubes and what effects they elicit.

For simplicity, it is usually assumed that a current is produced in some circuit of a tube because the respective electrode accepts some of the electrons travelling inside the envelope. The stream of these electrons constitutes what is called the *convection current*. In-depth studies of vacuum tubes have revealed that the current in the external circuit of any tube electrode is an *electrostatically induced current*. What is actually involved can be explained if we recall the process of *electrostatic induction*.

Let there be an uncharged conductor A (Fig. 19-3) one end of which is brought to the negatively charged end of another conductor B.

As a result, some electrons of conductor A, repelled by the charge on conductor B, will move to the opposite end of conductor A, giving rise to a negative image *charge*. The end nearest to the inducing charge on conductor B will experience a deficit of electrons or, which is the same, there will appear a *positive image charge* at that end. As this happens, a current flows along conductor A—this is an electrostatically induced current. Its magnitude increases with an increase in the inducing charge on conductor B and in the rate of approach towards conductor A. If we withdraw conductor B away from conductor A, the electrons will move back and, in consequence, a reverse current will flow in conductor A, of a magnitude again determined by the rate of travel of conductor B and by the inducing charge at its end.

To sum up, *when an electric charge is brought closer to or moved away from a conductor, an electrostatically induced current is produced in the conductor.*

In vacuum tubes, the role of an inducing negative charge is played by the electron stream, that is, the convection current. Inside a tube, this current always gives rise to electrostatically induced currents in the wires connected to the tube electrodes. The electrostatically induced current increases with an increase in the number and energy of moving electrons and also with a decrease in the spacing between the electron stream and a given electrode.

Suppose, as an example, that a direct voltage is applied to the anode of a thermionic diode. It causes a stream of electrons to move inside the tube from cathode to anode and to give rise to an electrostatically induced current in the external part of the anode circuit. Thus, the anode current begins to flow when electrons just move away from the cathode and not when they have reached the anode.

In the static or quasi-static conditions when $t_t \ll T$, the electrostatically induced current in the anode circuit is equal to the convection current. For this reason, it is legitimate not to invoke the concept of electrostatically induced current under these conditions. At microwave frequencies, however, when the electrode voltages change appreciably during the transit time, there is a need to consider the currents electrostatically induced in the various electrode circuits. In fact, it may even so happen that because of their inertia electrons will be oscillating in, say, the cathode-anode space without ever having any chance to reach the anode. Still, they will produce an electrostatically induced current in the anode circuit.

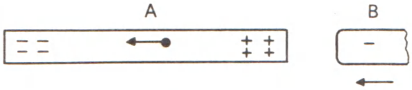


Fig. 19-3
Electrostatic induction

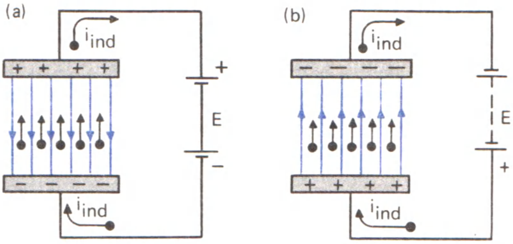


Fig. 19-4
Current induced by the motion of electrons in a d.c. field

The concept of electrostatically induced current provides better insight into the conversion of energy that takes place as electrons move through an electric field. As an example, let us discuss the motion of electrons in an accelerating or a retarding field between two electrodes, assuming that this field is set up by a battery (Fig. 19-4). In the battery circuit, the stream of electrons moving inside the tube produces an electrostatically induced current flowing in the same direction as the convection current. Here, as elsewhere, the arrow shows the direction of electron motion from the "−" to the "+" terminal, and not the conventional direction which is the other way around. It is easy to see that in an accelerating field (Fig. 19-4*a*) the electrostatically induced current flowing through the battery will be a discharge current for the battery. The battery is discharging, that is, expending its energy which is conveyed by the moving electrons and enhances their kinetic energy. On the contrary, in a retarding field (Fig. 19-4*b*) the electrostatically induced current will be a charging current for the battery, that is, electrons will give up their energy and it will be stored in the battery. Of course, the charging and discharging of a battery by an electrostatically induced current is not utilized for any practical purposes in microwave circuitry, and we have mentioned the process only as an illustrative example.

What is important, however, is the generation of electrostatically induced currents in the resonant circuits connected to a tube. Figure 19-5 shows a resonant circuit made up of an inductance L and a capacitance C which may be the capacitance between any two electrodes of the tube. Suppose that free oscillations exist in the resonant circuit. Owing to them, an alternating voltage can be detected at the resonant-circuit terminals and the tube electrodes. Also suppose that there is a stream of electrons moving between the electrodes (it is immaterial for the time being how it has been produced).

If the field set up by the voltage at the electrodes tends to retard electrons (Fig. 19-5*a*), the electrostatically induced current will be feeding the resonant circuit. The point is that the direction of flow of this current is such that the voltage it generates across the resonant circuit is in phase with the voltage existing in the resonant circuit owing to the free oscillations. This means that the electrostatically induced current opposes the decay of oscillations. In other words, some of the kinetic energy carried by the moving electrons is transferred to the resonant circuit and maintains oscillations there.

If, however, the field set up by an alternating voltage is an accelerating one for the electrons moving inside the tube (Fig. 19-5*b*), the electrostatically induced current will produce across the resonant circuit a voltage drop which is opposite in phase to the alternating voltage of free oscillations and it will speed up their decay. Now the resonant circuit expends some of its energy to accelerate the electrons, and so the decay of oscillations in the resonant circuit is speeded up.

To sum up, *if we wish to defer the decay of oscillations (that is, if we wish to maintain oscillations) in a resonant circuit connected to the electrodes of a tube, an electron stream should be injected into the interelectrode space at instants when the electric field is a retarding one.*

To refine our idea of the electrostatically induced current, let us see how it is produced in a diode. Our findings will be valid for any other system of two electrodes. To simplify the discussion, let us consider the case where the anode voltage is a rectangular pulse whose duration is comparable with the transit time of an electron. The waveforms of the voltage pulse and of the current electrostatically induced in the anode and cathode leads of the diode are shown in Fig. 19-6. The same figure shows the distribution of the electron stream, that is, the convection current, in the anode-cathode space at different instants.

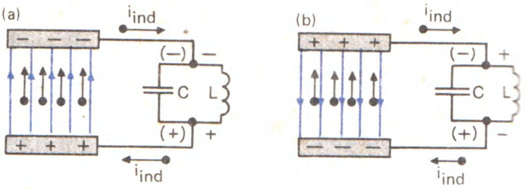


Fig. 19-5

Current induced by the motion of electrons in the field set up by the a.c. voltage of a resonant circuit

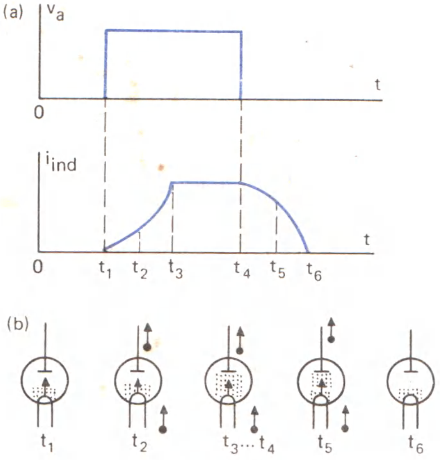


Fig. 19-6

Electrostatically induced current in a diode

At time t_1 the electrons just start moving away from the cathode (to be more exact, they start moving from the electron cloud near the cathode), and there appears an electrostatically induced current. The anode-cathode space is not yet filled with electrons, but at a later time, t_2, this space already contains a sizeable number of electrons. Since they are moving in an accelerating field, they have a higher velocity than they did at t_1. Owing to this, the electrostatically induced current builds up and the rate of build-up is increasing. At time t_3 the electrons reach the anode, and all of the anode-cathode space is filled with moving electrons. The electrostatically induced current is now a maximum. This situation is maintained until the voltage pulse ceases (at time t_4), following which no more electrons are moving from cathode to anode. In the meantime, the electrons filling the anode-cathode space keep moving towards the anode by inertia. Their number is decreasing—electrons are swept out of the anode-cathode space and the electrostatically induced current decreases in proportion (at time t_5). At time t_6, when all of the electrons have been swept out of the anode-cathode space, the electrostatically induced current is zero. As is seen, the electrostatically induced current pulse is stretched in time as compared with the voltage pulse and lags behind it—that is, it reaches its maximum value and decays to zero at later instants.

If a positive pulse of anode voltage is followed by a negative pulse, some of the electrons will reach the anode, but the remaining ones will be retarded so much that they will come to a stop and travel back to the cathode. In consequence, there will appear a reverse convection current and a reverse electrostatically induced current pulse. Similar events take place when an alternating sinewave voltage is applied to a diode.

19-4 The Input Resistance of and Power Dissipation in Tubes

A factor of special importance for any amplifier stage is the input resistance of the tube, that is, the opposition seen by the source of the signal being amplified.

In the generalized amplifier stage shown in Fig. 19-7 the signal source *SS* which generates an emf *E* and has an internal resistance R_{ss} is loaded into the input impedance of the tube. This impedance has an active (resistive) component and a reactive (capacitive) component. We are solely concerned with its resistive component, so it is denoted as R_{in}.

It is always desirable that R_{in} be as large as possible. Ideally, R_{in} should be infinity—then the grid circuit would be open, and the grid would draw no current. In consequence, there would be no voltage drop across the internal resistance of the signal source, and all of its emf would be impressed on the grid ($V_g = E$). In the circumstances, the signal source might have any power, however small. For R_{in} to be infinity, it is essential that electrons should not be intercepted by the grid and produce any grid current. In other words, the grid bias E_g must exceed the peak value of the alternating voltage (signal) being amplified: $|E_g| \geqslant V_{mg}$. In practice, a nearly ideal case occurs when the stage is operating at sufficiently low frequencies and we may neglect the capacitive current passing through the input capacitance of the tube.

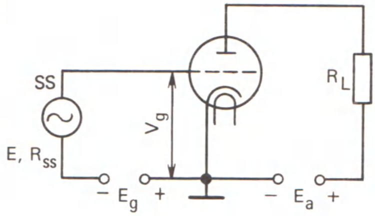


Fig. 19-7

An amplifier stage

At high frequencies, R_{in} is anything but infinity. As it increases, there is a proportionate increase in the alternating grid current I_g. Any increase in this current entails an increase in the voltage drop across the internal resistance of the signal generator, R_{ss}, and a decrease in the useful grid voltage because

$$V_g = E - I_g R_{ss}$$

More power is lost in the input resistance:

$$P_{in} = I_g^2 R_{in}$$

and the signal source has to deliver a greater total power.

It is convenient to describe the performance of an amplifier stage in terms of the stage voltage gain k_V which shows how many times the input voltage is boosted. At high frequencies, it is also important to know the stage power gain k_P defined as the ratio between the output and input power:

$$k_P = P_{out}/P_{in} \tag{19-4}$$

where P_{out} is the useful power delivered by the tube.

With a low input resistance, P_{in} may rise so high that k_P will fall to unity or even less. Obviously, it is not warranted to use amplifiers whose power gain is less than 2 or 3. At microwave frequencies, the input impedance abruptly decreases, and the power gain is small or even nil. The reduction in the input resistance of a tube at microwave frequencies is traceable to the electrostatically induced current that appears in the grid circuit.

The events taking place in a triode may take different paths, depending on the relative values of transit time and period of oscillation, the seperation between the cathode and grid and also between the grid and anode, and the electrode voltages. Whatever the situation, however, the electron-inertia effects observed at microwave frequencies will always give rise to a heavy electrostatically induced current in the grid circuit, leading to a drastic fall in the input resistance of the tube. For better insight into the matter, let us take a closer look at what happens in a triode in some special case.

Let the grid be fed with an alternating voltage in the form of rectangular positive pulses and a cutoff bias voltage (Fig. 19-8*a*). In the circumstances, the grid will remain negative at all times and so no electrons can reach it. Let the transit time of electrons from cathode to grid, t_{gk}, be equal to the transit time from anode to grid, t_{ag}, and slightly shorter than half the pulse duration. The waveforms of the electrostatically induced currents in the triode circuits for these conditions are shown in Fig. 19-8*b* and *c*, and the distribution of the convection current (that is, the electron stream) at different instants, in Fig. 19-8*d*. Suppose that the grid is wound with a pitch so short that the cathode-grid space and the anode-grid space may be treated as separate diodes.

Prior to time t_1, the tube is at cutoff, and there is no current flowing. At time t_1, the tube is cut in (that is, rendered conducting), and electrons begin to move from the cathode (to be more accurate, they begin moving from the electron cloud near the cathode) towards the grid, and the electrostatically induced current i_1 in the grid wire begins to build up. A similar current, i_k, equal to i_1, appears in the cathode wire as well. If at time t_2 the cathode-grid space is filled half-full with electrons, i_1 will be equal to some average value. From that instant on, it will keep rising until it becomes a maximum at time t_3 when the electron stream reaches the grid. Because the grid repels electrons, they pass through it and keep on moving towards the anode. This stream of electrons moving away from the grid induce in the grid wire a current, i_2, opposite in direction to i_1. In the anode wire, too, a current i_a is induced, equal to i_2.

At time t_4 the rising current i_2 has some average value and is a maximum at time t_5 when the grid-anode space is filled full with moving electrons. Until time t_6, the two currents i_1 and i_2 remain constant and equal to each other. At time t_6 the tube is driven to cutoff, and no electrons leave the cathode any longer. However, the electron stream filling the spacings between the electrodes keeps on moving.

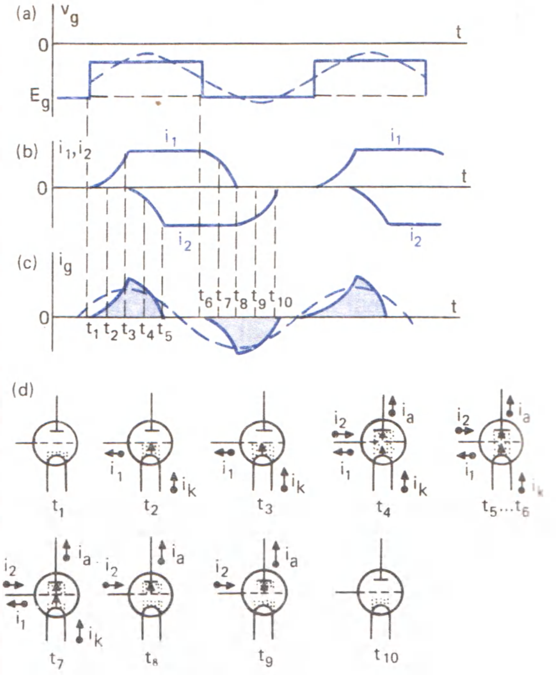


Fig. 19-8

Electrostatically induced current in the grid circuit of a tube

In the cathode-grid space, the electrons keep moving by inertia and approach the grid. Their number in this space is decreasing and the current i_1 is decreasing too. At time t_7 it has some average value, and at time t_8 it falls to zero because all of the electrons have been swept out of the cathode-grid space. Following that, the number of electrons in the grid-plate space decreases, and so does the current i_2. It falls to some average value at time t_9, while at time t_{10} when all the electrons have been collected by the anode it falls to zero.

Thus, two electrostatically induced current pulses opposite in direction are produced in the grid circuit (Fig. 19-8*b*). The resultant electrostatically induced grid current (Fig. 19-8*c*) is an alternating current. The dashed curves in Figs. 19-8*a* and *c* represent the fundamentals of voltage and grid current. As is seen, the fundamental of the electrostatically induced grid current leads the fundamental of voltage somewhat in phase. This means that the electrostatically induced grid current has both an active and a reactive (capacitive) component. The latter is lumped with the ordinary capacitive current which exists in the grid circuit due to the input capacitance of the tube. Because the total capacitive current is increased, electron inertia may be said to increase the input capacitance to some degree.

The worst consequence of electron inertia is the active (resistive) component of grid current. It is responsible for the fact that the input resistance falls with rising frequency, and this results in a reduced power gain. The input resistance is indicative of the power lost by the signal source connected in the grid circuit. This power is transferred by the active component of the electrostatically induced current from the signal source via the electric field to the electrons which acquire a greater kinetic energy and expend it to heat the anode. When the tube is operating at lower frequencies and the transit time of electrons may be neglected, the grid voltage shown in Fig. 19-8*a* gives rise to currents i_1 and i_2 which have the same rectangular waveform and the same duration. Because these two currents are equal in magnitude but opposite in sign, the net grid current is nil. In consequence, no power supplied by the signal source is wasted in such a case.

We have examined the effect of electron inertia on the tube performance when the grid is fed with positive rectangular voltage pulses only approximately. However, the general trend remains the same in the more elaborate cases: an active induced current is produced in the grid circuit or, to state this differently, one consequence of electron inertia is that the tube comes by a resistive input impedance. If the alternating voltage applied to the grid consists of both positive and negative pulses, the latter set up a retarding field which drives some electrons back to the cathode. They are accelerated by the field, and some of the energy supplied by the signal source is dissipated as heat when the electrons bombard the cathode.

The picture is far more complex when the alternating voltage is sinusoidal in waveform. But again, in operation at microwave frequencies a resistive electrostatically induced current is brought about in the grid circuit at the expense of some of the energy supplied by the signal source. In the final analysis, this energy is dissipated as heat at the anode and cathode due to the action of the convection current (the electron stream). The point is that the positive half-cycles of grid voltage accelerate the electrons leaving the cathode to higher energies while during the negative half-cycles the grid repels the electrons moving towards the anode, and they too acquire more energy. As a result, the electrons bombard the anode more vigorously and raise it to a higher temperature. The electrons that fail to pass through the grid and fall back to the cathode are likewise repelled by the grid during the negative half-cycles of grid voltage and acquire still more energy. These electrons bombard the cathode and heat it more. Thus, the signal source transfers energy to electrons over the entire cycle of grid voltage, and the electrons waste it in bombarding the anode and cathode.

We have given only a rough sketch of what actually happens in a tube at microwave frequencies, but it gives an ample idea about the events. A more rigorous analysis of tube performance at microwave frequencies is too complicated and is beyond the scope of this text.

Theory gives us the following equation for the resultant electrostatically induced grid current I_g in the case of an alternating grid voltage V_g:

$$I_g = kg_m f^2 t_{gk}^2 V_g \qquad (19\text{-}5)$$

where k = proportionality factor taking care of the electrode structure and supply voltages

g_m = transconductance of the tube

t_{gk} = transit time from cathode to grid

Hence, the input resistance is given by

$$R_{in} = V_g/I_g = 1/kg_m f^2 t_{gk}^2 \qquad (19\text{-}6)$$

The values of k, g_m and t_{gk} remain constant for a given tube and a given set of supply voltages. On replacing them with a single factor and passing from frequencies to wavelengths, we get

$$R_{in} = a\lambda^2 \qquad (19\text{-}7)$$

The factor a is difficult to calculate, and the calculated results are heavily in error. Therefore, it is customary to find it experimentally for many tubes. Thus found, it takes care not only of electron inertia but also of other processes causing loss of power. For some receiving-amplifying tubes operating at normal supply voltages the factor a is several hundreds. If $a = 400\ \Omega\ \text{m}^{-2}$ and $\lambda = 50$ cm, then $R_{in} = 400 \times 0.5^2 = 100\ \Omega$.

As is seen, the input resistance is very small, and this might lead to a prohibitive reduction in gain. The point is that for a stage using a pentode the stage voltage gain is

$$k_V \approx g_m R_L \qquad (19\text{-}8)$$

where R_L is the anode load resistance. If the load is a resonant circuit, it would be placed in shunt with the input resistance of the tube in the next stage, so the overall load resistance would be less than R_{in}. For an amplifier operating at a wavelength of 50 cm and using tubes for which $g_m = 5$ mA/V and $R_{in} = 100\ \Omega$, it may be taken approximately that $R_L = 100\ \Omega$. Then, $k_V \approx \approx 5 \times 10^{-3} \times 10^2 = 0.5$. Thus, instead of gain, the stage would produce an attenuation.

The calculation of R_{in} given above holds only for low alternating voltages. At high alternating voltages (such as generated by oscillators and transmitters), a more elaborate procedure would be necessary.

The loss of power in tubes operating at microwave frequencies may be caused by other factors as well. For example, *skin effect* is responsible for an increased resistance of the electrodes and leads. Heavy currents flow over the surface of the metal conductors and these are wastefully heated. More power is likewise lost in the solid dielectrics exposed to an alternating electric field, say, in the glass envelope.

Heavy losses of power in tubes impair the efficiency of microwave amplifiers and oscillators, cause the tubes to be overheated, and drastically reduce the Q-factor of the associated resonant circuits. Resonant circuits in the form of coaxial transmission lines or cavity resonators have a Q of several thousands or even several tens of thousands. However, when they are connected to a tube, their Q is drastically degraded (often by an order of magnitude or even more). This is equivalent to the impairment in the Q of a conventional resonant circuit when it is shunted by a relatively small resistance.

19-5 Tubes in Pulse Operation

Vacuum tubes used in microwave transmitters are often operated in the pulsed mode. For example, nearly all radar transmitters generate pulses with a duration of units or tens of microseconds spaced apart by time intervals of a greater duration (Fig. 19-9). In this mode of operation, the average power is a small fraction of the pulse power. Suppose, for example, that the pulse duration is $\tau_p = 10$ µs, the pulse power is $P_p = 100$ kW, and the pulse repetition frequency is $f = 200$ Hz. Hence the pulse repetition period is

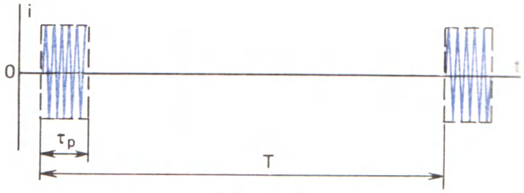


Fig. 19-9

Microwave oscillations in pulsed operation

$$T = 1/200 = 0.005 \text{ s} = 5000 \text{ µs}$$

or 500 times the pulse duration. Therefore, the average power is one-five hundredth of the pulse power:

$$P_{av} = 0.2 \text{ kW}$$

An important indicator of pulse-circuit performance is the *pulse duty factor* defined as the ratio of the average pulse duration to the average pulse repetition period. (The pulse repetition period is often termed the *pulse spacing* defined as the time interval from one pulse to the next, that is, between the corresponding times of two consecutive pulses. The term ‘pulse interval’ is ambiguous — it may be taken to mean the duration of a pulse instead of the space or interval between pulses. One should also be guarded against an indiscriminate use of the term ‘pulse separation’ — the interval between the trailing edge of a pulse and the leading edge of the succeeding pulse.)

In Soviet practice, it is more customary to use the reciprocal of the pulse duty factor, that is, the *pulse period-to-pulse duration ratio*

$$Q = T/\tau_p \qquad (19\text{-}9)$$

Hence,

$$P_{av} = P_p/Q = P_p\tau_p/T \qquad (19\text{-}10)$$

Tubes for pulse operation usually have anodes of a small surface area because the anode dissipation is decided by the average power. High-power pulses are produced by feeding very high voltages to the grid and anode during a very short time. For example, the anode voltage may be as high as tens of kilovolts. To avoid a breakdown, special care is taken to make the insulation between the electrodes and their leads as perfect as practicable, and also to maintain a high vacuum.

The cathode of a tube used for pulse operation must be able to maintain a very high level of electron emission. This purpose can well be served by oxide cathodes whose emission in pulse operation is tens of times the figure obtainable in continuous working. Under usual condition, an oxide cathode has an emissivity of up to 0.5 A cm^{-2}, and the cathode efficiency is 100 mA W^{-1}. In pulse working, the emissivity of an oxide cathode runs as high as 70 A cm^{-2} and the cathode efficiency is raised to 10 000 mA W^{-1}.

The high emissivity in pulse operation is due to the ejection of a great number of electrons from the oxide coating under the action of the strong external electrostatic field which penetrates into this semiconductor coating. This level of emission can be maintained by an oxide cathode only if the pulse duration does not exceed 15-20 μs and the pulse spacing is of a longer duration. An attempt to maintain a high emissivity for a longer time might lead to the 'poisoning' of the oxide cathode, with the result that the emission current would rapidly fall off. The high level of emissivity could then be restored only after the cathode has been given an ample time for 'rest'.

Apart from oxide cathodes, tubes for pulse operation can use the bariated-tungsten cathode (or the L cathode, after its inventor H. J. Lemmens), the thoriated-oxide cathode, cermet cathodes made of a mixture of thorium and powdered molybdenum, etc. The emissivity of some of them in pulse working is as high as 300 A cm^{-2}.

19-6 Microwave Vacuum Tubes

Microwave vacuum tubes are built so as to minimize their interelectrode capacitances and lead inductances and the electrode spacings. Measures are also taken to minimize power losses. Among other things, the envelopes are fabricated from special grades of low-dielectric-loss glass or r. f. ceramic materials. In the case of transmitting tubes, special emphasis is placed on the cooling of the anode and of the tube as a whole because the anode dissipation is high and the device might otherwise be overheated.

UHF tubes can of course operate at lower frequencies, but for use in the SHF band they are unsuitable. Some of the bantam and miniature glass-button tubes can be used in oscillators and amplifiers operating in the UHF band (at frequencies of several hundred megahertz).

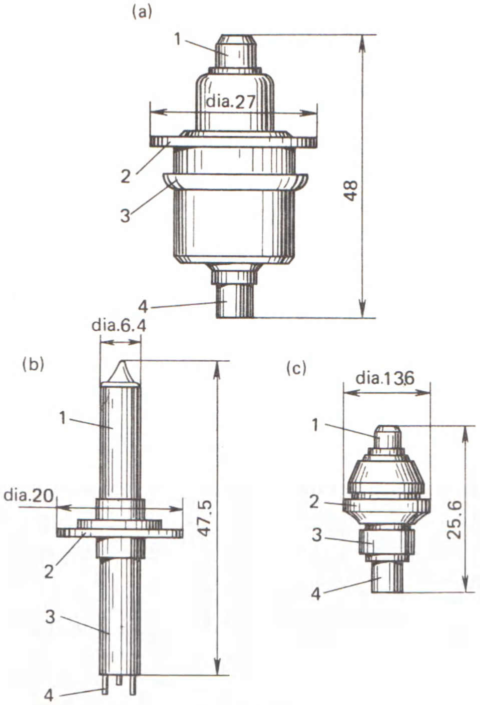


Fig. 19-10

Microwave triodes: (*a*) metal-glass; (*b*) pencil; (*c*) miniature metal-ceramic. (*1*) Anode connection; (*2*) grid connection; (*3*) cathode and heater connection; (*4*) heater connection

Disc-seal and cylindrical-terminal tubes have specially been designed for operation in the UHF band and at the lower end of the SHF band. The cylindrical and disc seals varying in diameter provide for connection of a tube to coaxial resonant lines or cavity resonators and act as parts of the respective tuned circuits or systems.

An example of such tubes is the metal-glass triode shown in Fig. 19-10*a*. In this tube, one of the connections is made common to the heater and the cathode disc seal. The tube can be used as an oscillator at frequencies as high as 3.6 GHz and delivers 0.1 W of useful power. This design is employed in some special-purpose diodes.

Special mention should be made of the *pencil* or *cryon triode* (Fig. 19-10*b*). This is a metal tube with cylindrical anode and cathode connections and a disc-seal grid. It can operate as an oscillator or an r. f. amplifier at frequencies up to

3 GHz and deliver as much as 5 W of useful power. Pencil or cryon diodes and triodes are also available in other designs.

Of considerable interest is the subminiature triode using cylindrical connections (Fig. 19-10*c*). It is intended primarily for use in the input stages of common-grid amplifiers in microwave receivers. At a maximum frequency of 3 GHz this type of triode has a 12-fold gain in power and a 40-fold power amplification at 1.2 GHz.

Some of the tubes in the metal-ceramic series are able to operate at frequencies as high as 10 GHz. Nuvistors, small vacuum tubes with a cantilever-supported cylindrical electrode that eliminates the need for mica supports, can likewise operate in the UHF band. Only metal and ceramics are used, and there is no getter in a nuvistor.

Larger oscillators, r.f. amplifiers and transmitters, notably those used for pulse work, employ metal-ceramic transmitting triodes similar in design to the receiving-amplifying tubes examined earlier. They are likewise adapted for connection to coaxial resonant systems. Figure 19-11 shows both the external appearance and the internal arrangement of a metal-ceramic transmitting tube. The cathode, grid and anode surfaces are discs spaced closely apart. Sometimes, the electrodes are dished. The contact from the indirectly heated oxide cathode is a cylinder which serves at the same time as one of the heater contacts. The other end of the heater has a contact enclosed inside the cylinder. The grid contact is a cylinder which is part of the tube envelope.

The anode is a substantial cylinder with its top bonded to a ceramic cylinder acting as part of the envelope. At the opposite end this cylinder is bonded to the grid seal. The grid and cathode connections are separated by a ceramic disc. These metal-to-ceramic seals or bonds are the special feature of metal-ceramic tubes. These tubes use a special grade of ceramic materials which suffer low power losses at microwave frequencies. The anode is cooled by means of a finned radiator screwed onto the anode stud. The radiator is blown over with air supplied by a fan. Such tubes may well operate without a radiator, but at a substantially lower anode dissipation and power output.

Metal-ceramic tubes may be designed for CW duty and for pulse work.

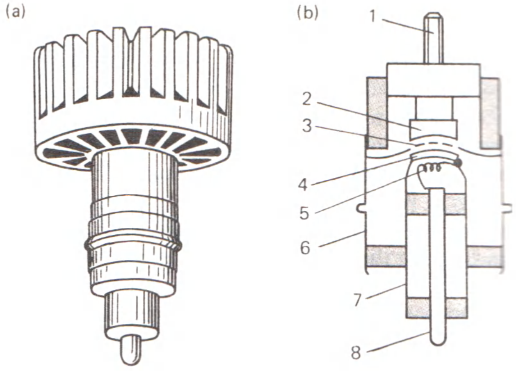


Fig. 19-11

External appearance and arrangement of a metal-ceramic transmitting triode: (*1*) anode radiator support stud; (*2*) anode; (*3*) grid; (*4*) cathode; (*5*) heater; (*6*) grid connection; (*7*) cathode and heater connection; (*8*) heater connection

Tubes more elaborate in their electrode structure than triodes are seldom used in the UHF band. The point is that with several grids enclosed in the same envelope a greater spacing has to be provided between the anode and the cathode, and this would lead to a longer transit time. In receiving tubes, an increase in the number of electrodes causes a build-up in tube noise. All in all, UHF oscillators, r.f. and power amplifiers mostly use triodes. Recently, however, newer types of tetrodes have been developed for use at these frequencies. One of them is a metal-ceramic beam tetrode for frequencies up to 1 GHz with an output of 2 kW. There are also dual beam tetrodes for ultra-high frequencies.

It is customary to set up triode amplifier stages into a common-grid circuit so as to avoid the generation of unwanted oscillations due to parasitic coupling via the interelectrode capacitances (Fig. 19-12). A special feature about this circuit configuration is that the input *LC*-circuit is placed in the cathode wire. The control grid is returned to the chassis ground at one end and connected to the "–" side of the anode supply at the other. In this arrangement, the grid doubles as a screen grid and minimizes spurious coupling between the anode and grid circuits via the anode-cathode interelectrode capacitance C_{ak} rather than the coupling via the anode-grid capacitance C_{ag} as is the case in conventional common-cathode amplifier stages. To serve as a good screen, the grid is wound with a fine pitch, and so these triodes have a large amplification factor (100 and higher). Owing to the closely wound grid, the anode-cathode interelectrode capacitance is reduced to a few hundredths of a picofarad.

A limitation of the common-grid configuration is that it has a low input resistance. This is because the input current in such a circuit is the cathode current, while in a common-cathode circuit the input current is substantially smaller because it is the grid current. Practically, the input resistance for a common-grid circuit is approximately equal to $1/g_m$. If a tube has a transconductance of 5 mA/V, then $R_{\text{in}} = 1/5 = 0.2$ kΩ. The signal source is loaded by a small R_{in} and has to supply a good deal of power. Despite this limitation, the common-grid configuration is employed often because it operates consistently and does not jump into unwanted self-excited oscillations.

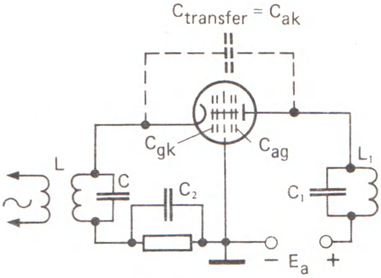


Fig. 19-12

Common-grid amplifier stage

Chapter Twenty

Advanced Types of Microwave Tubes

20-1 General

There is quite a number of advanced types of tubes for use at microwave frequencies. They depend for their operation on the fact that the electrons acquire their kinetic energy from the electrostatic field set up by a power supply and transfer some of this energy to a microwave electromagnetic field which retards the electrons. They are in fact *wave-interaction tubes.*

There are two broad classes of *wave-interaction* (or *wave-type*) tubes. They are *O-type* (or *linear-beam*) tubes and *M-type* (or *crossed-field*) tubes. In O-type devices a d.c. magnetic field is either absent or used solely to focus the electron beam. M-type tubes use crossed E and H fields, that is, a d.c. electric field at right angles to a d.c. magnetic field, which is the reason why they are called crossed-field wave-type tubes. The joint action of the two fields governs to a large extent the path travelled by the electrons.

Historically, the first representatives of O-type tubes have been *klystrons* widely used even today. Rather than being a nuisance, the transit-time effects in the klystron are essential for normal operation. The basic types of klystrons are the *drift-tube two-cavity* (or *multicavity*) *klystrons* and the *reflex klystrons.* The former are good as oscillators, r.f. and power amplifiers, and the latter only as oscillators. The O-type class also includes the *travelling-wave tube* (TWT) and the *backward-wave tube* (BWT). It is important to add, however, that there are also *M-type TWTs* and *M-type BWTs.* Historically, the first M-type tube has been the *maagnetron.* Recently, more M-type tubes have been developed, such as the *Amplitron* (a broadband crossed-field amplifier tube with a re-entrant electron stream; the electron stream interacts with the backward wave of a non-reentrant r.f. structure) developed by the Raytheon Manufacturing Company (USA), and the *Stabilotron* (which is actually the same as the Amplitron but it is called the Stabilotron when used as a self-excited stabilized oscillator) also developed by Raytheon.

In more detail the wave-type tubes are discussed in the sections that follow.

20-2 The Drift-Tube Klystron

In sketch form, the arrangement and connection of a two-cavity drift-tube klystron are

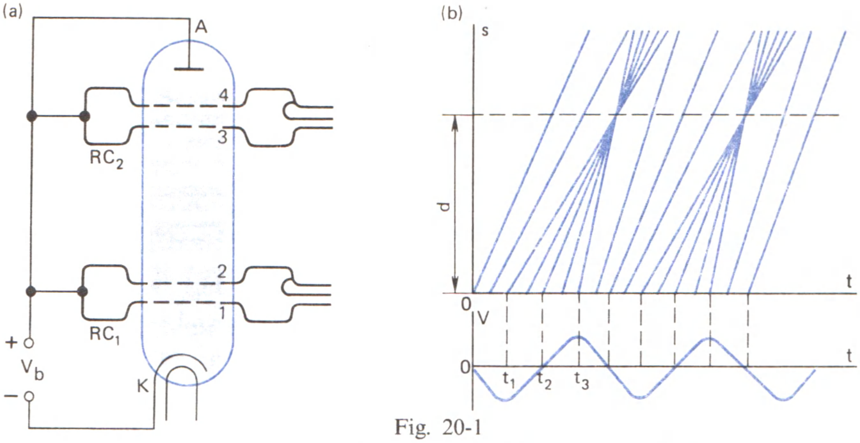


Fig. 20-1

Structure and operation of a two-cavity drift-tube klystron

shown in Fig. 20-1*a*. The tube is primarily intended for use as an amplifier. The electron stream emitted by the cathode of an electron gun moves via two pairs of grids towards a collector electrode. The grids are the walls of two resonant cavities, RC_1 and RC_2 (sometimes, the grids are just openings made in the cavities). RC_1 is the *input cavity* to which the signal to be amplified is fed at frequency f over a coaxial line and a coupling loop. Its grids *1* and *2* are called the *buncher grids* because they make up what is known as the *buncher*. Owing to its action, a small amount of drive power fed to the input cavity causes a high a.c. voltage to appear across the cavity gap formed by the grids. The resulting velocity modulation and bunching action produce a large a.c. current in the beam. RC_2 is the *output cavity* which is located a suitable distance from the first and a high voltage is induced across its two grids, *3* and *4*, called the *catcher grids*. Useful power is coupled out of the output (catcher) cavity by means of a coupling loop and a coaxial line.

The accelerating field that causes electrons to enter the buncher cavity at a high velocity u_0 is produced between the buncher grid *1* and the cathode by a positive potential V_b applied to both cavities and the collector.

When oscillations take place in the buncher (input) cavity, RC_1, an alternating electric field exists between the buncher grids *1* and *2*. This field alternately accelerates and decelerates the incoming electron stream during the consecutive half-periods of the r.f. cycle. During the half-cycle when grid *2* is positive and grid *1* is negative, the field between them is accelerating, and the electrons moving through the buncher gap acquire an additional velocity Δu. During the next half-cycle when grid *2* is negative and grid *1* is positive, the field becomes decelerating, and the velocity of the passing electrons is diminished by Δu. Only the electrons that pass through the buncher at the instant when the applied voltage is zero are able to travel at u_0.

Thus, the electrons finding themselves between the catcher grids *2* and *3*, called the *drift space*, have all different velocities. There is no r.f. electric field in the drift space because there is no potential difference between grids *2* and *3*, and the electrons keep on moving at velocities that remain unchanged. Here, the electrons that were accelerated in the input cavity during one half-cycle catch up with those that were decelerated during the previous half-cycle. As a result, the electron beam is divided into individual bunches having a higher current density. It may be said that owing to the velocity modulation of the electron stream, this stream is density modulated in the drift space.

The way an electron bunch is formed can be illustrated graphically. Figure 20-1*b* relates the distance *s* travelled to the time *t* expended by electrons moving through the buncher at different instants, and also variations in the r.f. voltage across the input cavity. The distance *s* is reckoned from the buncher gap. The electrons move through the drift space at a uniform velocity, so their travel can be represented by straight lines whose slope is indicative of the travel velocity.

Consider the motion of three electrons moving through the buncher at times t_1, t_2, and t_3. Let the electrons enter the buncher at the same velocity and also let their transit time through the buncher be shorter than the period T. Then the electron passing through the buncher at time t_2 keeps on moving at the same velocity u_0, and its travel is represented by a straight line having some average slope. The line representing the travel of the electron passing through the buncher at time t_1 has a smaller slope because this electron was decelerated in the cavity and it has a lower velocity. In contrast, the electron passing through the buncher at time t_3 is accelerated in the cavity, and the line representing its travel runs steeper. The three lines intersect at about the same point. This means that the three electrons are bunched together at this point along their path. Any other electrons passing through the buncher at the intermediate instants will arrive at this point likewise at about the same instant. The electrons moving through the buncher before t_1 and after t_3 will not be bunched together, and this fact is illustrated in the plot.

Thus, a bunch can only be formed by the electrons passing through the buncher during the same half-cycle of the r.f. field. A good bunching action is possible only when the velocity modulation depth of electrons is moderate, that is, when the changes in the velocity of the electrons under the influence of the modulating r.f. field are insignificant in comparison with the velocity they acquire from the d.c. accelerating voltage. Therefore the r.f. voltage across the cavity gap must be a fraction of the d.c. voltage V_b. The bunching action is repeated during one half-cycle of each period.

The d.c. voltage V_b is chosen such that the electrons are bunched in the catcher, that is, at a distance d from the buncher. If V_b is too high, electrons will be bunched at a greater distance (between the catcher and the collector). If V_b is too low, this will happen too close to the buncher (in the drift space). It follows therefore that V_b must be maintained at a precisely defined and stabilized value.

Past the point of the maximum bunching action, the electrons spread apart again. If we extend the plots of electron travel, we will see that the bunching action again occurs first at a distance $3d$, then at a distance $5d$, etc. This action is not, however, utilized for practical purposes because there is no advantage to be gained from increasing the size (mostly, the length) of the klystron.

To repeat, the catcher receives electron bunches that follow one another at frequency f. They give rise to induced current pulses and excite oscillations in the output cavity. For these oscillations to reach a maximum amplitude, the output cavity must be tuned to the same frequency f to which the input cavity is tuned. Much as the anode current pulses in an r.f. amplifier stage traverse the anode resonant circuit and build up amplified oscillations there, so in the klystron the electron stream made up of bunches gives rise to amplified oscillations in the output cavity. The amplification occurs due to the energy supplied by the source of V_b setting up the accelerating field. The electrons acquire more energy in this field and, since they are velocity-modulated in the input cavity, they give up some of this energy to the output cavity.

Electron bunches, or *current-density waves*, pass through the output cavity when its electric field produces a decelerating effect. The electrons passing through the output cavity are intercepted by the collector electrode and heat it. Some electrons are also intercepted by the buncher and catcher grids. If the electron beam were not velocity-modulated, it would not be able to maintain oscillations in the output cavity because a uniform electron stream would absorb energy from the cavity during the half-cycles when its field is accelerating and give up the same amount of energy during the half-cycles when its field is decelerating. As a result, no energy would be transferred from the electrons to the cavity.

The same reasoning can be applied to the interaction between the electron stream and the input cavity. If a uniform electron stream entered this cavity, it would take away some energy from the cavity during one half-cycle and return the same amount of energy during the next half-cycle. No energy would be withdrawn from the cavity over a whole cycle, so no velocity modulation of the electron beam would take place.

We have not considered electron inertia, however. Although they travel through the buncher very fast, the electrons do have some inertia, and because of it some energy is expended to effect modulation. To minimize it, it is customary to raise V_b and to minimize the spacing between the buncher grids. Because little power is lost in the input cavity, its input impedance and Q-factor are fairly high.

Two-cavity klystrons can yield a power gain of tens. Their major limitation is that their efficiency (defined as the ratio between the useful oscillatory power in the output cavity to the power supplied by the d. c. collector source) is not more than 20% although the theoretical limit is as high as 58%. Failure to achieve this theoretical figure can be explained by several factors. For one thing, the current density produced by the bunching action is not high enough because the electrons are emitted by the cathode at different velocities. Therefore, when they are passing through the buncher at the same instant of time, they differ in velocity. For another, there is mutual repulsion between the electrons in the stream. Furthermore, the inertia of the electrons passing through the catcher is responsible for the fact that some oscillatory energy existing in the output cavity is taken away by electrons. Last but not least, some electrons are not bunched at all, and this means that they do not contribute to the useful effect. All in all, a sizeable proportion of the total energy is wasted to heat the grids and collector because in the long run all of the electrons reach these electrodes at some velocity.

Two-cavity klystrons are used as amplifiers in microwave transmitters. Their useful power in the CW mode may be tens of kilowatts. In pulse working, the figure is tens of megawatts. Power output decreases with decreasing wavelength (with increasing frequency).

Since the bunching action produces large-amplitude harmonics, the klystron can be used as a frequency multiplier if the output cavity is tuned to a higher harmonic. Multiplication ratios of 10 to 20 are typically used.

Klystrons are a poor choice as amplifiers for weak signals because of the strong inherent noise.

Increased efficiency results when one or more resonant cavities are placed between the input and output cavities. This is the reason why, at this writing, preference is given to *multicavity klystrons*. Although they are more elaborate in design, they offer a number of advantages over two-cavity devices.

In a multicavity klystron, the first cavity is input and the last cavity is output. The intermediate cavities are only connected to the "+" terminal of the power supply (Fig. 20-2). The pulsating electron stream produces oscillations and, as a result, an alternating electric field which additionally modulates the electron beam, and enhances the bunching action. In consequence, higher current-density waves reach the output cavity, and the efficiency and power gain of the device are greatly increased.

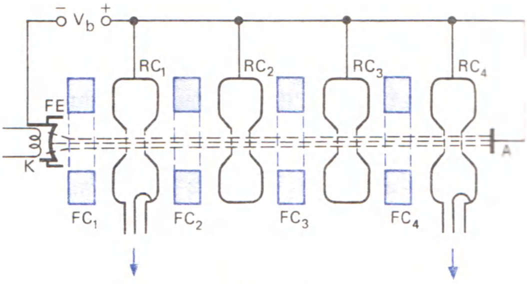


Fig. 20-2

Structure of a multicavity drift-tube klystron: *FC*, focusing coils; *FE*, focusing electrode

State-of-the-art drift-tube klystrons widely differ in output power, number and type of cavities, methods of beam focusing, manner of coupling r. f. energy in and out of the cavities, frequency tuning, cooling, and some other features. They may be designed for either continuous-wave (CW) or pulsed work.

In pulsed klystrons, the pulse repetition rate is usually from tens to thousands of hertz and the pulse duration from a fraction of a microsecond to milliseconds. In terms of power, drift-tube klystrons may be classed into low-power, medium-power, high-power, and extra-high-power. The respective pulse power is less than 10 kW, from 10 kW to 1 MW, from 1 to 100 MW, and over 100 MW. The power ratings of CW klystrons are by a factor of 10^3 lower. It may be added that we have quoted figures applicable to klystrons for use in the UHF band. In the SHF band, the respective ratings are lower.

Beam focusing may be electrostatic, electromagnetic (by a coil), or magnetic (by permanent magnets). Energy can be coupled in and out of the cavities by means of coaxial, waveguide, or combination (coaxial-waveguide) devices. The cavities may be internal, that is, built into the klystron itself, or external. Most often klystrons are fabricated tuned to some fixed frequency, but there are also *tunable klystrons* in which the cavities can be tuned to any desired frequency mechanically. Unfortunately, *mechanical tuning* is a complicated procedure and enables the tuning frequency to be changed by not more than 10-15%. High-power klystrons may be naturally or forced (water or air) cooled.

Multicavity klystrons may be 40-50% efficient but for most of them the figure is markedly lower. Their power gain is sometimes tens of thousands. Practically, a power gain of more than 10^6 is difficult to obtain. High-power klystrons, especially those for pulse work, need supply voltages of tens or even hundreds of kilovolts.

Drift-tube klystrons have a very narrow bandwidth because there are several tuned cavities. As a rule, the bandwidth does not exceed several megahertz. The overall bandwidth may somewhat be expanded by staggered tuning of the cavities, but this will lead to a lower gain. Power output may sometimes be increased by using multi-beam klystrons. In them, several electron beams pass through the same cavities in parallel.

The drift-tube klystron can be made to oscillate (that is, converted to a self-sustained oscillator) by providing a feedback loop from the output to the input cavity in the form of a coaxial cable. The length of the cable is chosen such that the wave fed back is in phase with the oscillations in the input cavity. Then the current-density waves (electron bunches) will pass through the output cavity during the half-cycles when the field is a decelerating one, and maintain the oscillations. If they were fed back in anti-phase, they would withdraw energy from the output cavity, and the oscillations would die out rapidly. Sometimes, the input and output cavities of a two-cavity klystron have a common wall. In such a case, diffraction feedback is utilized by making a hole in the common wall.

Drift-tube klystrons are used as self-sustained oscillators very seldom. In fact, a far better performance for local oscillators (usually in the low-power rating class) is obtained with the reflex klystron which has only one cavity and so there is no need for a precise tuning of several cavities as in a multicavity klystron.

20-3 The Reflex Klystron

The connection and arrangement of the *reflex* (*single-cavity*) *klystron* are shown in Fig. 20-3. Here a *reflector*, or *repeller*, located past the cavity resonator turns the beam back upon itself, so that the cavity acts as both the buncher and the catcher. The electron beams are accelerated by a high d.c. voltage V_b, and the reflector is maintained at a voltage V_r negative with respect to the cathode. For better beam focusing the cathode is enclosed in a cylinder which is called the *focusing electrode* and is usually connected to the cathode. Energy is coupled out of the resonator with the aid of a coupling loop and a coaxial line.

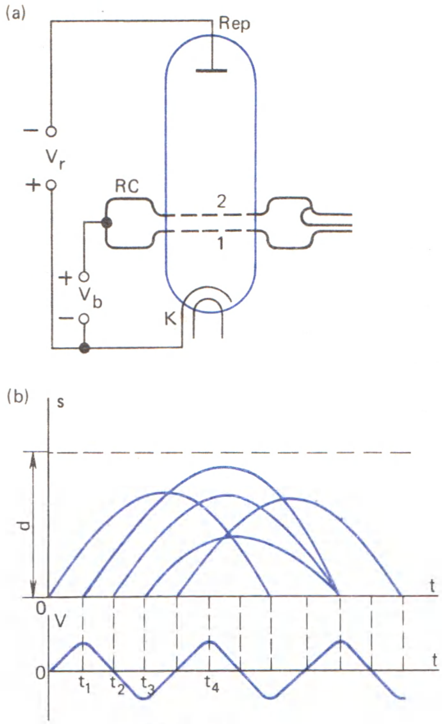


Fig. 20-3

Arrangement and operation of a reflex klystron

Driven by the accelerating field, the electron beam enters the cavity and produces an electrostatically induced current pulse in it. This gives rise to oscillations which excite an alternating electric field between the walls of the cavity. This field velocity-modulates the electron beam. As a result, the electrons leave the cavity at different velocities and enter the *drift* (or *interaction*) *space* between the resonator and the repeller where a d.c. decelerating field is at work. The electrons entering this field are slowed down, brought to a stop, and finally accelerated backwards into the resonator. The higher the velocity of an electron, the deeper it moves in the decelerating field and the longer it stays there. As a result, the electrons passing through the resonator during the positive half-cycles and accelerated more by the alternating electric field can return the same instant as do the electrons passing through the resonator at a later time (that is, during the negative half-cycles) and decelerated by the alternating field.

This can conveniently be illustrated by the following analogy. If we throw upwards three identical objects, but the first at a maximum velocity and the third with a minimum velocity, the three may fall back all at the same time. The first will rise farther than the remaining two and will be in motion for a longer time, and the last will rise least of all and will return the shortest time interval later.

Although velocity modulation in the reflex klystron occurs in the same way as it does in the drift-tube klystron, the bunching action is different. Figure 20-3 shows the paths of electrons in the reflex klystron to explain the principle of the bunching action. These paths are not straight lines, but curves (parabolas) because each electron is moving nonuniformly. At first, an electron is slowed down (as far as the stop point), then it is accelerated back upon itself. The electrons leaving at times t_1, t_2 and t_3 come back all at the same instant or, which is the same, there appears a local current-density wave or an electron bunch. The same reasoning applies to the electrons passing through the resonator at the intermediate instants between t_1 and t_3.

A current-density wave may return into the resonator at different instants, depending on the values of V_b and V_r. On returning, the current-density waves give up their energy to the resonator only when they enter a decelerating field, that is, when grid *1* is negative and grid *2* is positive (for the forward-travelling electron beam this will be an accelerating field). The electrons give up most of their energy when they return at the instants of a maximum decelerating field in the resonator. When they return at any other instants, they give up less energy, and the oscillations carry less power. Indeed, the oscillations may cease altogether if the returning electrons supply too little energy. If a current-density wave returns during the negative half-cycles when there is an accelerating field in the resonator, the electrons tap energy from the resonator, and the oscillations decay at a faster rate.

It is customary to state the transit time in the drift space, t_t, that is, the time interval between the instant when electrons leave the resonator in the forward direction and the instant when they return to the resonator, for some central or axial electron (ejected at time t_2) around which the remaining electrons are grouped. In Fig. 20-3 this time is 1.75 *T*. By raising the magnitude of the negative voltage at the repeller we can cause the electron bunch to return at time t_4, that is, at the instants spaced 0.75 *T* apart. Conversely, by lowering the repeller voltage, we can cause the electrons to travel farther into the retarding field and to return to the resonator at the instant 2.75 *T* later. At all of these instants, the electron bunches give up the largest amount of energy to the resonator because they enter the retarding field when it is at its maximum intensity. Thus, if we wish to maintain undamped oscillations carrying the largest amount of power in a reflex klystron, we must satisfy the following condition:

$$t_t = (n + 3/4)\,T$$

or (20-1)

$$t_t = (n + 3/4)/f$$

where *n* is any integer, including zero, called the *mode number*. Hence, several modes can exist in the reflex klystron. When $n = 0$ and $t_t = 0.75\,T$, we have the zeroth mode. When $n = 1$ and $t_t = 1.75\,T$, we have the first mode. The second mode corresponds to $n = 2$ and $t_t = 2.75\,T$, etc. Figure 20-4 shows the paths of electron bunches for the first three modes.

Several factors affect the transit time of electrons in the reflex klystron. An increase in the separation *d* between the resonator and the repeller leads to a decrease in the strength of the retarding field at the same potential difference $V_b - V_r$. A weaker field, however, retards the electrons less, and they move farther into the field and return at a later time. Thus, given a large value of *d*, we may obtain a higher mode number.

The transit time is strongly affected by the repeller voltage – this fact is illustrated in Fig. 20-4. An increase in the magnitude of V_r leads to a stronger retarding field

$$E = (V_b - V_r)/d$$

and the klystron operates in a mode of a lower

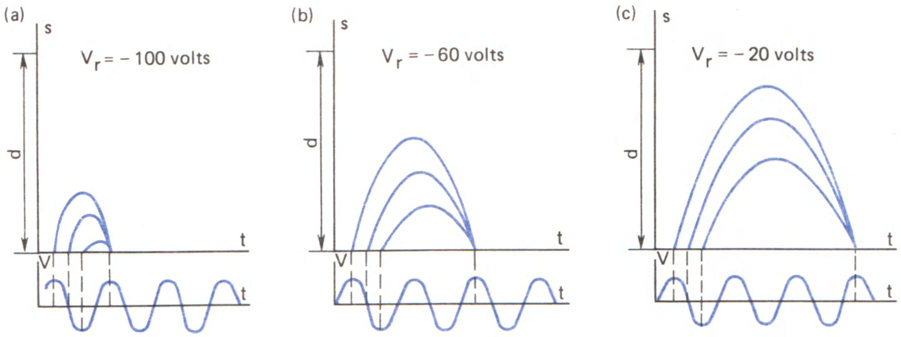


Fig. 20-4

Motion of electrons in an operating reflex klystron in (*a*) the zeroth voltage mode, (*b*) the 1st voltage mode, and (*c*) the 2nd voltage mode

number. The way variations in V_r affect the power of oscillations in the resonator is demonstrated in Fig. 20-5. This power is usually a maximum for some single mode where the bunching action is most effective (the current density is at its largest). For modes of the higher or lower numbers the power of oscillations is smaller because of the poorer bunching action. This impairment in bunching may be caused by the mutual repulsion of the electrons in the beam, the difference in initial velocity between the emitted electrons, variations in field strength between the drift space and the grids, and some other factors.

The d.c. beam-accelerating voltage at the resonator, V_b, has a weaker influence on the transit time. Its variations lead to two opposing effects which balance each other to some extent. If, for example, we raise V_b, the electrons will gain in speed and tend to move farther into the drift space or, which is the same, their transit time will tend to increase. A rise in V_b, however, leads to a rise in the strength of the retarding field in the drift space so that the electrons are slowed down more and tend therefore to return at an earlier time or, which is the same, the transit time tends to be shorter.

If we move to an ever greater mode number by progressively reducing the magnitude of the negative voltage at the repeller, we will in the long run find ourselves in a situation where at $V_r > 0$ the electrons will be intercepted by the repeller and fail to return to the resonator.

Reflex klystrons are not more than 3-5% efficient; sometimes they may be even less than 1% efficient. Therefore they are manufactured for power outputs not more than 1 W. Most often reflex klystrons are used as local oscillators for receivers and test equipment. Their available power output is usually several hundredths or tenths of a watt.

Reflex klystrons can be tuned in several ways. One is capacitive tuning which consists in that the separation and, as a consequence, the capacitance between the resonator grids is changed by some mechanical means. This type of tuning is ordinarily employed in reflex klystrons with built-in resonators. The tuning range is usually 5-10%. In the case of an external resonator, the tuning frequency can be varied by as much as 20% by moving a metal plunger inside the resonator. Tuning must be accompanied by adjustment of, say, the repeller voltage so as to assure the best conditions for self-sustained oscillations. In such cases, it is said that *the repeller voltage must be tracked with the cavity tuning.*

Narrow tuning ranges can be obtained by varying the repeller voltage alone. This is *electronic tuning*. If we raise the magnitude of the negative repeller voltage, the electron bunches

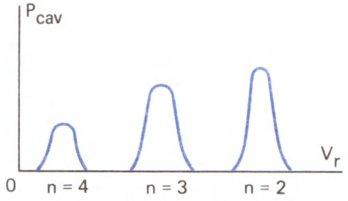


Fig. 20-5

Power output of a klystron as a function of the repeller voltage will return to the cavity at an earlier time and the frequency of oscillations will be increased. If we lower the magnitude of the negative repeller voltage, the electron bunches will return to the cavity at a later time, and the frequency of oscillations will be decreased. There is a mechanical analogy for electronic tuning. Suppose that a pendulum is kept swinging by pushing it from without. If we push the pendulum when it is in one of its extreme positions, the frequency of oscillations will be equal to the natural frequency of the pendulum. However, we might push the pendulum a bit earlier, that is, before it reaches its extreme positions. As a result, the frequency of oscillations would somewhat rise. If we wish to lower the frequency of oscillations, we should push the pendulum at a later time during each swing.

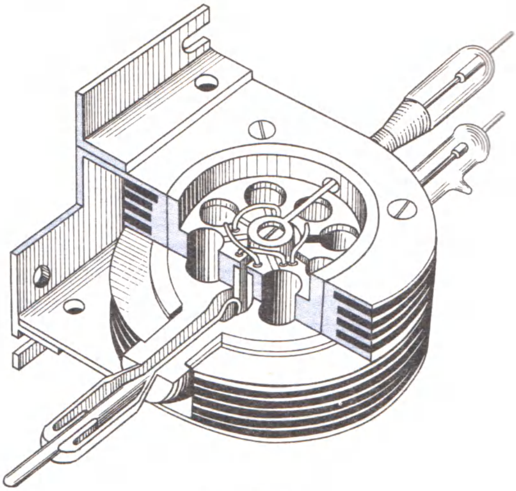


Fig. 20-6
Structure of a magnetron

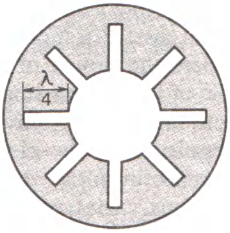


Fig. 20-7
Magnetron cavities as quarter-wavelength lines

Electronic cavity tuning leads to a fall in power output. Therefore, it is ordinarily used so that power output is not more than halved. The electronic tuning range is as a rule several tens of megahertz on either side of the central frequency. One volt of change in repeller voltage leads to a frequency change of a few tenths of one per cent of the operating frequency, that is, several megahertz. In specially designed klystrons, the electronic tuning range may be as wide as 10-15%.

The fact that the repeller voltage has a strong effect on the power output and signal frequency offers a means for amplitude, frequency, and pulse modulation effected by applying the modulating voltage to the repeller.

Because reflex klystrons fall solely in the low power class, the accelerating voltage taken from the supply source is usually 250-450 V, being as high as 2500 V only in some special klystrons. The beam current may be tens of milliamperes.

The strong effect of supply voltages (notably the repeller voltage) on the signal frequency necessitates the use of regulated (stabilized) power supplies in many cases.

20-4 The Magnetron

The *magnetron* is an extremely important type of high-power microwave oscillator. It is used in radar transmitters, particle accelerators, microwave-heating units, etc. The underlying principle of the magnetron is the interaction between an electric and a magnetic field, on the one hand, and electron beams, on the other, leading to the generation of oscillations at microwave frequencies. At present, the most commonly used form of device is the *multicavity magnetron.*

The structure of a multicavity magnetron is shown in Fig. 20-6. It is a high-vacuum diode containing a cathode and an anode. The cathode is mostly of the indirectly heated, oxide-coated large-area type. At the ends the cathode has discs which prevent electrons from moving axially. The anode is a substantial copper block usually divided into two or more segments. The evacuated space between the cathode and anode is called the *interaction* space which is in communication via slots left between the segments of the anode so that an even number (say, 8) of *resonant cavities* are formed in the form of cylindrical holes. Each slot acts as a capacitor. Alternating electric charges are formed on its surfaces, and an electric field is set up in each slot. The cavity inductance is supplied by the cylindrical surface of each cavity or hole equivalent to one turn of an inductor. The large surface area of each 'turn' serves to reduce both its resistance and inductance. Such a resonator stands somewhere between a lumped-constant resonant circuit and a quarter-wavelength resonant transmission line. What we have described is sometimes called a *hole-and-slot magnetron*. In some devices, resonators take the form of slots alone, and quite aptly one then has a *slot magnetron*. The slots are quarter-wave long (Fig. 20-7).

There is a very tight coupling between the magnetron cavities because the alternating magnetic field of one cavity links the adjacent cavities (Fig. 20-8). Also, there are *straps* which are wires or strips connecting the ends of the segments in the anode as shown in Fig. 20-6.

For better cooling, the outer surface of the anode has the form of a finned radiator sometimes blown over with air. Together with the copper discs bonded to the anode, the structure forms an envelope which is then evacuated and maintained at the desired vacuum. The heater leads are passed inside glass tubes sealed to the anode. The cathode is usually connected to one of the heater leads.

Microwave energy is extracted from the magnetron by means of a coupling loop placed in one of the cavities and connected to a coaxial line. The coaxial line is brought out likewise inside a glass tube. Owing to the tight coupling between the cavities, microwave energy is extracted from all the resonators. At very short wavelengths, microwave energy is extracted by means of a waveguide coupled to a resonator by a slot. Sometimes both a coaxial line and a waveguide may be used to extract the output microwave power.

The anode of a magnetron is highly positive with respect to the cathode. Since the anode also doubles as the envelope of the magnetron, it is usually grounded while the cathode is held at a high negative potential. This produces between the anode and cathode an accelerating field whose lines of force run radially as in a diode with cylindrical electrodes. Parallel with the axis of the magnetron is a strong d. c. magnetic field set up by a magnet whose pole pieces enclose the magnetron. One likely arrangement of the magnet system is shown in Fig. 20-9. In what are called *packaged magnetrons*, the permanent magnets are made integral with the tube itself.

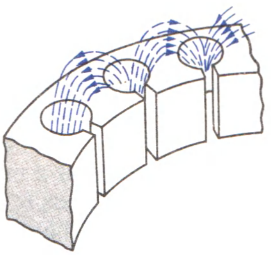


Fig. 20-8

Magnetic coupling between adjacent anode cavities (resonators)

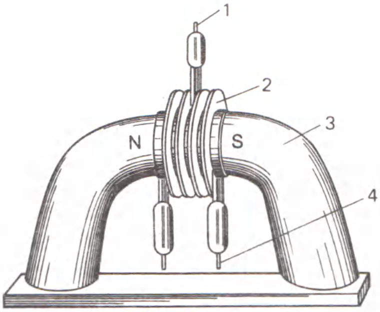


Fig. 20-9

Magnetron using an external magnetic system: *1*, microwave connection; *2*, radiator; *3*, magnet; *4*, heater connection

Let us trace the motion of electrons in a magnetron, assuming that no oscillations exist in its cavities. To simplify matters, the anode is shown unslotted (Fig. 20-10). Influenced by an accelerating electric field, the electrons tend to follow the lines of force, that is, to travel radially, towards the anode. However, as soon as they come up to a certain velocity, the constant magnetic field which is at right angles to the electric field bends their paths. The radius of curvature gradually increases with increasing velocity of the electrons. For this reason, the electrons follow paths which are high-order curves. The figure shows the travel paths of an electron leaving the cathode at a negligibly low initial velocity for several values of the magnetic flux density (or magnetic induction) B, with the anode voltage held at a constant value.

At $B = 0$, the electron travels along radius *1*. At a magnetic flux density below some critical value, B_{cr}, the electron travels to the anode along a curved path *2*. At B_{cr}, a more curved path *3* results. Now the electron moves near the anode surface without touching it, and goes back to the cathode. Finally, at a magnetic flux density above B_{cr}, the electron makes a still tighter U-turn somewhere between the anode and the cathode (curve *4*) and returns to the cathode.

Magnetrons are operated at a magnetic flux density which somewhat exceeds B_{cr}. Therefore, in the absence of oscillations, the electrons move very close to the anode surface but at different velocities because they set out on their journeys at different initial velocities. Since a very great number of electrons are in motion, a rotating electron sheath is formed around the cathode (Fig. 20-11). Of course, the electrons do not remain in this rotating sheath permanently. The electrons leaving the cathode earlier return there to be replaced by the electrons leaving it at a later time. The rotating sheath moves at an angular velocity determined by the anode voltage – at a high anode voltage the electrons pass by the anode at a very high velocity. If no electrons are to be intercepted by the anode, the magnetic flux density must be made higher.

The rotating electron space charge produced by the joint action of the crossed d. c. electric and magnetic fields interacts with the alternating (r. f.) electric fields of the resonators and maintains oscillations in them. This interaction is an extremely complicated process, so we will discuss it in general terms only.

To begin with, let us see how oscillations are brought about in the cavities. Since all the cavities are tightly coupled to one another, they make up a composite resonant system which has several natural frequencies. When the electron beam just starts revolving past the cavity slots (such as when the anode voltage is turned on), electrostatically induced current pulses are produced in the cavities, and damped oscillations are elicited. They may differ widely in frequency and phase. For example, in the case of a symmetrical system, the oscillations initiated in the cavities will be in phase. Unfortunately, complete symmetry can never be achieved. So other oscillations are initiated, shifted in phase from one another.

The largest available power output and the highest efficiency are obtained when the oscillations in one cavity are in anti-phase (that is, 180° out of phase) with those in the other cavities – this is known as *π-mode operation.* Figure 20-12 shows the r. f. electric fields for this mode of operation, the polarities of the r. f. potentials at the anode segments, and the directions of the currents flowing over the surfaces of the cavity holes. Because we know the role played by the d. c. electric field that accelerates the electrons and imparts them the necessary kinetic energy, the field is not shown in the figure for simplicity.

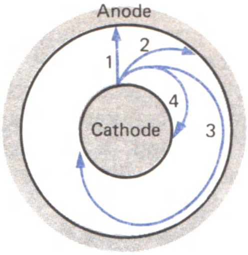


Fig. 20-10

Effect of a magnetic field on the motion of electrons in a magnetron

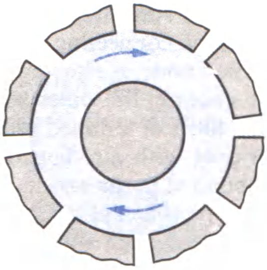


Fig. 20-11

Rotating electron space-charge sheath in a magnetron in the absence of oscillations

In π-mode operation there is a very tight inductive coupling between the cavities because the magnetic flux from one moves into another (see Fig. 20-8). As a rule, magnetrons are operated in the π-mode, and measures are usually taken to facilitate the onset of these oscillations. This purpose can, for example, be served by placing straps between alternate anode segments on which the potential is of the same polarity. Other modes will then die out rapidly.

The electrons interact with the r. f. electric field in such a way that, with the operating conditions properly adjusted, the electron beam transfers to the field more energy than it withdraws from it. This is essential for the oscillations initiated in the cavities to become undamped. There are several factors that facilitate

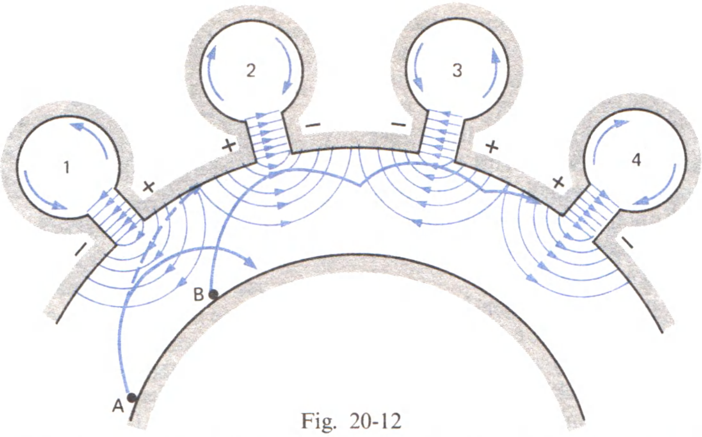


Fig. 20-12

Travel paths of an unfavourable electron (*A*) and of a favourable electron (*B*) in a magnetron in the presence of oscillations in the anode cavities (resonators)

energy transfer from the electron stream to the cavities.

Above all, the r.f. electric field sorts out the electrons into *'favourable'* and *'unfavourable'* ones, with the latter swept out of the interaction space and driven back to the cathode. Let us take a closer look at this action.

For the electrons travelling clockwise (Fig. 20-12), the odd-numbered cavities (*1*, *3*, etc.) produce accelerating electric fields, while the even-numbered cavities (*2*, *4*, etc.) produce decelerating fields. In fact, every half-cycle an accelerating field becomes decelerating and vice versa. The figure shows the travel paths of two electrons. Electron *A* enters an accelerating field and absorbs some energy from the cavity – it is an unfavourable electron moving away from the cavity slot and returning to the cathode. If only a d.c. field were present, this electron would follow the path shown dashed. However, the r.f. field of cavity *1* bends the path followed by the electron and adds more energy to it so that it goes back to the cathode. Unfavourable electrons bombard the cathode and produce what is known as *back heating* – a factor that has to be reckoned with when using magnetrons. It reduces the heater current required when the tube is running and also stimulates secondary electron emission. Therefore, the cathode surface must be made strong enough to avoid its destruction by electron bombardment.

Electron *B* entering the retarding r.f. field set up by cavity *2* follows a more intricate path. This electron contributes some of its energy to the cavity, and what is left is not enough for the electron to return to the cathode. It loses all of its energy at some point in the interaction space before it reaches the cathode, then it is again accelerated towards the anode and its path is at the same time bent by the magnetic field.

In a magnetron with properly adjusted anode voltage and magnetic flux density, the time it takes a favourable electron to travel from one slot to the next is equal to a half-period. On arriving at the slot of cavity *3*, this electron finds itself in a retarding field again because what was an accelerating field a half-period before now becomes a retarding field. In consequence, the electron again contributes some of its energy to the cavity and follows a still shorter path in the direction of the cathode. Finally, on having expended a sizeable proportion of its energy, the electron is collected by the anode. Of course, what we have described is only an approximation to the true path travelled by a favourable electron.

Favourable electrons contribute to the cavities more energy than unfavourable electrons absorb from them. The point is that an unfavourable electron absorbs energy only from one cavity and also it moves past the slot at a fairly large distance from it, that is, in a weak r.f. field. So it can absorb a very small amount of energy from the cavity. In contrast, a favourable electron contributes energy to two cavities and travels closer to their slots, that is, in a stronger r.f. field.

Energy transfer from the electrons to the cavities is promoted by electron-stream modulation similar to the modulation taking place in a two-cavity klystron. Every preceding cavity in a magnetron acts as a buncher for the rotating sheath of electrons, and every next cavity as a catcher. However, the modulation pattern here is more complicated. In a two-cavity klystron, the advancing electron stream is velocity-modulated so that a bunched beam is produced. The bunching takes place in the drift space free from electric and magnetic field.

In a magnetron, the rotating sheath of electrons is likewise influenced by the r. f. alternating electric field of a given cavity and is velocity-modulated. However, this field is anything but uniform, so it changes not only the velocity but also the travel path of the electrons. The process is complicated by the fact that it takes place in a radial d. c. electric field which affects the velocity of electrons and, jointly with the d. c. magnetic field, their trajectories.

An outcome of this velocity modulation and path bending is that the rotating sheath of electrons is bunched into a configuration resembling spokes of a wheel (Fig. 20-13). The number of electron 'spokes' is half the number of resonant cavities. Of course, these 'spokes' are diffused and not sharply defined.

When the operating conditions of a magnetron are properly adjusted, the space-charge cloud rotates at an angular velocity such that the 'spokes' move past the slots at the instants when a retarding r. f. field exists there. In contrast, the spacings between the 'spokes' move past the slots at the instants when an accelerating field exists. As a result, the electron sheath transfers some of its energy to the cavities and also loses some of its energy as the electrons back-heat the cathode and heat the anode. All of this energy comes from the anode supply.

The number of cavities N, the magnetic flux density B, and the frequency of oscillations f are connected by a relation of the form

$$NB = af \qquad (20\text{-}2)$$

where a is a coefficient dependent on magnetron design.

The magnetic flux density is connected to the anode voltage by the following equation:

$$B = bV_a^{1/2} \qquad (20\text{-}3)$$

where b is a constant.

It is seen from Eqs. (20-2) and (20-3) that for operation at higher frequencies a magnetron needs a greater number of resonators or a higher magnetic flux density and a higher anode voltage.

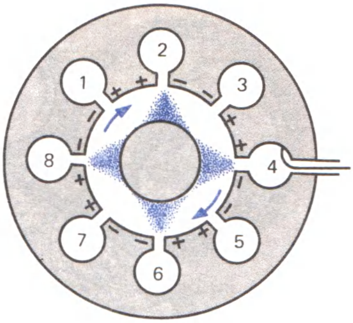


Fig. 20-13

Rotating electron space charge sheath in a magnetron in the presence of oscillations in the anode cavities (resonators)

As a rule, the magnetic flux density is anywhere between 0.1 and 0.5 tesla. Magnetrons intended for pulsed work in the UHF band are built for a power output of tens of megawatts, and those for the SHF band, for a power output of several megawatts. The pulse anode voltage of the largest magnetrons may be as high as tens of kilovolts and the anode current, several hundred amperes. Magnetrons for CW operation are able to deliver an output power of tens of kilowatts in the UHF band and several kilowatts in the SHF band. High-power magnetrons use forced air or water cooling. They may be 70% or even more efficient when operated in the UHF band. In the SHF band, the efficiency falls to 30-60%.

Most magnetrons are fixed-frequency units but *tunable magnetrons* have also been produced. In them, the tuning frequency is varied by varying the natural frequency of the resonators. This is done by inserting copper cylinders into the resonator holes in order to reduce the inductance and thus to raise the tuning frequency, or by inserting metal sheets in order to increase the conductance and thus to lower the tuning frequency. The tuning range in such cases is not more than 10-15%. Also, difficulties arise in implementing these tuning mechanisms because they have to be placed inside an evacuated envelope and operated from the outside.

The electronic tuning of magnetrons is based on the fact that their frequency is a function of their anode current. A change of 1 A in anode current may cause the frequency of the tube to change by as much as several tens of megahertz. It is to be added, however, that this type of tuning has not found any appreciable use in conventional magnetrons.

The situation is different in the case of *voltage-tunable magnetrons* (one example is the *Mitron tube*) - by varying their anode voltage and, in consequence, their anode current, it is possible to change their tuning frequency by a factor of two (over an octave). These tubes somewhat differ from conventional magnetrons in design. One of their distinctions is that the anode current is limited by reducing the emission of electrons from the cathode (which is in turn achieved by operating the cathode at a reduced heater voltage). Another distinction is provision of an external resonant cavity of a low Q, that is, with a broad bandwidth. Voltage-tunable magnetrons built to operate in the continuous-wave (CW) mode and with an octave tuning range deliver an output power of several watts. With a tuning range of 5-20%, such tubes can deliver tens of watts of output power.

Conventional magnetrons are not stable enough in terms of frequency and phase. A greatly improved stability in π-mode operation can be achieved with coaxial magnetrons (Fig. 20-14). In a coaxial magnetron, there is a high-Q resonant cavity which encloses the anode block. This external resonant cavity has a natural frequency of its own, equal to the frequency of oscillation of a π-mode magnetron. It is coupled to the internal resonators by slots cut in the alternate segments of the anode block. With this arrangement, the oscillations set up in the resonators coupled to the external cavity are all in the same phase, and those existing in the alternate cavities are in anti-phase.

A far better choice for use at the higher end of the SHF band (that is at the shorter end of the centimetre-wave band) is offered by the *inverted coaxial magnetron*. In this type of magnetron, the cathode is a cylinder enclosing the anode block and electrons are emitted from its inner surface. In turn, the anode block encloses a high-Q resonant cavity intended to stabilize oscillations and coupled to the anode resonators.

A more recent addition to the magnetron family is the *nigotron* developed by P. L. Kapitsa of the Soviet Union. The nigotron is essentially a cylindrical cavity resonator with a d. c. magnetic field applied along its axis. The cavity encloses a coaxial cathode-anode assembly in which both the cathode and the anode are made in the form of segments. The high Q of the main cavity assures a high stability for the frequency of oscillations. When used in the UHF band and in the CW mode, the nigotron is able to deliver a power output of 100 kW or even more and is up to 50% efficient.

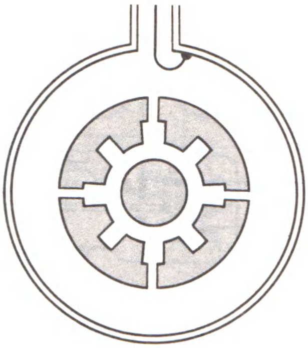


Fig. 20-14
Structure of a coaxial magnetron

20-5 Travelling-Wave and Backward-Wave Tubes

The drawbacks associated with the klystron amplifier (see Sec. 20-2) are to a great extent minimized in the *travelling-wave tube* (TWT) and the *backward-wave tube* (BWT). Also, the TWT can give a far higher gain and efficiency than the klystron because the electron beam in the TWT interacts with an alternating magnetic field over a longer path and is able to contribute more energy to generate amplified oscillations. The TWT uses a weaker electron beam and so its inherent noise level is relatively low. The bandwidth may be broad because the TWT proper does not contain any resonant system. The frequency span ratio (that is the ratio between the maximum and minimum tuning frequencies) may be from 2-to-1 to 4-to-1. The bandwidth is limited by the accessories coupling the tube to the external circuits rather than by the TWT itself. TWTs for use at frequencies of several gigahertz have a bandwidth of several hundred megahertz which is more than enough for radar applications and all forms of state-of-the-art telecommunications.

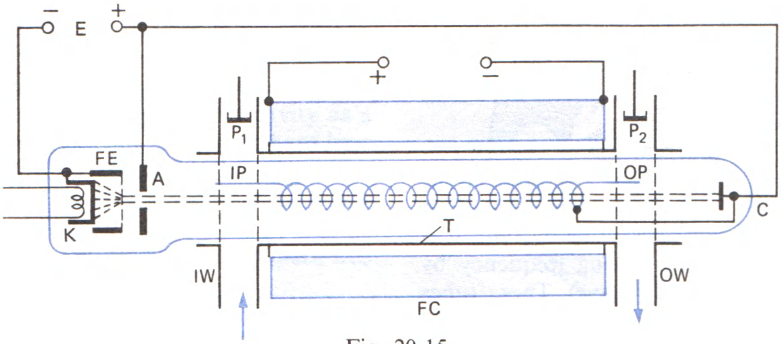


Fig. 20-15
Structure of an O-type TWT

In sketch form, the structure of an O-type (that is, linear-beam) TWT is shown in Fig. 20-15. The left-hand part of an extended envelope encloses an electron gun made up of an indirectly heated cathode K, a focusing electrode FE, and an anode A. The electron beam produced by the electron gun then travels through what is called the *slow-wave* (or *delay-line*) *structure* which may be made in the form of, say, a helix acting as the inner conductor of a coaxial line. The outer conductor of this line is the *body electrode* in the form of a metal tube, T. The helix is mounted on suitable insulators (not shown in the figure). The d.c.-energized focusing coil 'compresses' the electron beam all along its length, thereby preventing its spread due to the mutual repulsion of the beam electrons. Instead of a coil, the focusing action may be supplied by permanent magnets set up in a *magnetic focusing structure.* Unfortunately, magnetic focusing structures are cumbersome and compare unfavourably with the more recent *electrostatic focusing structures* which rely for their operation on an electrostatic field.

The signal to be amplified is fed to the TWT by an input waveguide IW enclosing an r.f. input pin IP which is the start of the helix. The helix terminates in an output pin OP which excites r.f. oscillations in the output waveguide OW. There are two pistons, P_1 and P_2, which are used to match the waveguides to the helix so that a travelling wave could propagate along the slow-wave structure. On emerging from the helix, the electron beam is picked up by a collector C which is electrically connected to the helix. In TWTs for frequencies up to 4 GHz the helix is coupled to external circuits by coaxial lines because waveguides that could be used at such frequencies would be too bulky.

The helix is proportioned so that the phase velocity of the wave propagating along the helix axis is

$$u_{\text{ph}} \approx 0.1c = 0.1 \times 300\,000 = 30\,000 \text{ km s}^{-1}$$

where c is the velocity of light in a vacuum. As a rule, the helix is wound with tens or even hundreds of turns. In TWTs for the SHF band the helix may be 10-30 cm long, and its diameter may be several millimetres.

Figure 20-16 shows the electric field pattern inside the helix for the case where the wavelength is equal to the distance spanned by six turns of the helix. A cross-sectional view of the helix is shown, with the "+" and "−" signs indicating the potential distribution on the helix conductor; here the bold signs correspond to higher potentials. The field pattern shown corresponds to one particular instant. Since the wave travels along the helix, the field rotates about and moves along its axis at the phase velocity u_{ph}. Of course, there is also an electric field between the helix and the outer metal tube (not shown in the figure), but it does not interact with the electron beam. The helix turns are surrounded by an r.f. alternating magnetic field, but it does not interact with the electrons either.

The electrons entering the helix must have a velocity slightly in excess of u_{ph} approximately equal to $0.1c$. This is achieved by setting the anode voltage at somewhat higher than 2.5 kV. As the electron beam interacts with the electric field of the travelling wave, the electrons are velocity-modulated and bunched. In this way, local beam-density variations are produced, with denser portions separated by more rarefied portions.

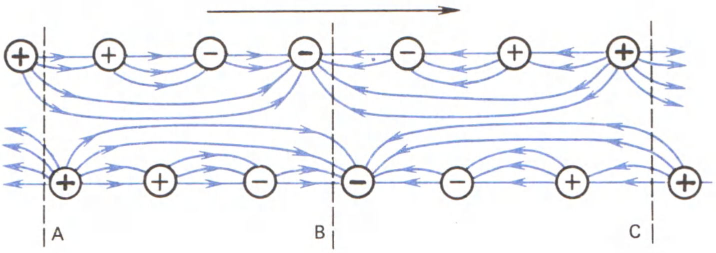


Fig. 20-16

The electric field due to a travelling wave inside the slow-wave structure

From reference to Fig. 20-16 it is an easy matter to see that portion *AB* of the helix (spanning a distance equal to a half-wavelength) produces a field which retards the electrons, while portion *BC* (spanning a distance equal to an adjacent half-wavelength) produces a field which accelerates the electrons. Thus, the field pattern existing along the helix is one of alternate accelerating and decelerating fields. If, at a given instant, a retarding field region exists at the start of the helix, the entering electrons will be slowed down and, on moving within this region, will be grouped into denser bunches. As they gradually slow down, they are continuously contributing energy to the field and beefing up the travelling wave. If, on the other hand, an accelerating field region exists at the start of the helix at a given instant, the entering electrons will pick up speed and, by overtaking the field, will enter the next region where a retarding field exists. Although these electrons absorb some energy from the travelling wave as they enter an accelerating field, they subsequently give it back to the wave because they move into a retarding field region.

To sum up, electron bunches are formed within the retarding field regions and contribute some of their energy to the wave continuously. The total energy contributed to the wave along the entire helix is quite substantial. As the travelling wave advances along the structure, its current and voltage rise in peak value towards the end of the helix. This is accompanied by a proportionate build-up in the strength of the accelerating and decelerating fields and in the bunching action. The aggregate result of the process is that greatly amplified oscillations emerge at the output. The energy contributed to the travelling wave is acquired by the electrons from the anode supply.

At high gain and at a less than perfect match with the waveguides, there appears a wave reflected from the output end of the helix. On reaching the input end, this wave is again reflected, amplified, then again reflected from the output end, and so on. In the long run, it results in the self-excitation of the tube which then generates oscillations of its own – a fact intolerable when the tube is intended for use as an amplifier. To avoid this, some of the helix – at the start or in the middle – is made of high-resistance wire which absorbs the energy carried by the reflected wave. As often as not, this purpose may additionally be served by a coat of graphite applied to the inside of the envelope or to the insulators that give support to the slow-wave structure.

TWTs intended for service at the higher end of the SHF band (that is, at the shorter wavelengths of the centimetre waveband) use other configurations for the slow-wave structure because the very small helices needed would be unprofitably difficult to make. Other slow-wave structures are also used at high power levels because no helix would be able to stand up to the resultant power dissipation. TWTs with slow-wave structures of other than the helix type are fabricated for power outputs up to 1 kW and for frequencies up to 10 GHz.

At this writing, quite a number of different TWTs are commercially available for use as input, intermediate and output broadband r.f. power amplifiers. Since the beam current contains harmonics, TWTs may well be used as frequency multipliers.

In terms of power output, TWTs may be classed as follows. Low-noise TWTs with a beam current of 100-200 μA can deliver several thousandths or hundredths of a watt. In special-purpose receivers the noise level can further

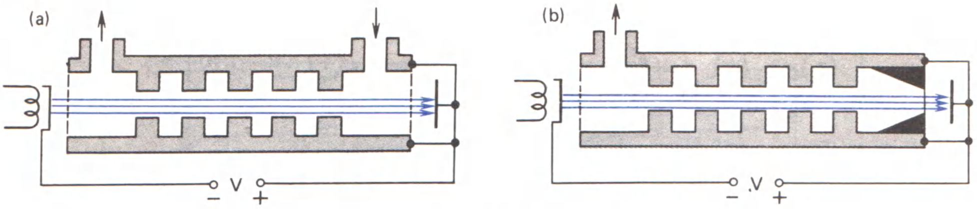


Fig. 20-17
Structure of (*a*) an O-type TWT amplifier and (*b*) an O-type TWT oscillator

be lowered by cooling the TWT to a very low temperature. Low-power TWTs (1-2 W) carry a beam current of units or tens of milliamperes and give a gain of several hundred thousands. Medium-power TWTs (up to 100 W) and high-power TWTs (up to 100 kW) have a gain of several thousands and carry a beam current from several hundred milliamperes to several amperes. Extra-high-power TWTs are able to deliver hundreds of kilowatts of useful power. The supply voltage ranges from several hundred volts for low-power TWTs to tens of kilovolts or even higher for high-power TWTs. High-power TWTs may be 30-40% efficient. Many TWTs are used in pulse work and are able to deliver a pulse power of 10 MW or even more.

The efficiency of a TWT can be enhanced by retarding the electrons as they emerge from the slow-wave structure. This is done by applying a lower d. c. voltage to the collector than to the slow-wave structure. In this way, less power is drawn from the power supply. Another way to enhance the efficiency is to utilize the bunching action based on the principle utilized in klystrons. An example of such tubes is the *Twystron.* This hybrid device consists of a klystron driving section and a TWT output section. In the Twystron, the klystron driving section produces electron bunches which then enter the TWT output section that delivers a highly amplified output signal. The Twystron is up to 50% efficient and the relative bandwidth is 10-15%. Pulse power output from some Twystrons is tens of megawatts.

The principle underlying the operation of travelling-wave tubes has served as the basis for the development of the backward-wave tube (BWT) some versions of which are called *Carcinotrons.* In contrast to TWTs, backward-wave tubes are mainly used as microwave oscillators, although some of them may operate as amplifiers as well. The BWT uses a focusing structure and a slow-wave structure similar to those used in the TWT, but it differs in that the electron beam travels in the opposite direction from the wave in the slow-wave structure and, as a result, produces the feedback required for oscillation.

In sketch form (that is, with its focusing structure omitted) an O-type amplifier BWT is shown in Fig. 20-17*a.* Its input is located at the collector and its output, at the cathode. Although this tube has no resonant system, it does possess resonant properties. The tube provides amplification only within a very narrow bandwidth, and the location of this band depends on the magnitude of d. c. accelerating voltage *V.* By varying it, we can accomplish electronic tuning.

A much wider use is made of O-type BWTs as oscillators (Fig. 20-17*b*). Instead of an input, a backward-wave oscillator has a lossy terminating section (the shaded triangles) at the collector. This section absorbs the wave reflected from the output end of the slow-wave structure. Such a wave might appear when the match at the output is less than perfect, so its absorption improves the performance of the backward-wave oscillator.

An O-type backward-wave oscillator operates as follows. Originally, oscillations exist in the tube due to the fluctuations of the electron beam. These variations are then amplified and the oscillations build up until they reach the desired magnitude. It is to be noted that oscillations may arise in a backward-wave amplifier as well if its beam current exceeds some critical value. The frequency of the oscillations generated by a backward oscillator depends on the accelerating voltage *V.* Therefore, it is possible to achieve electronic tuning ranges of 1.5 : 1 to 2 : 1. In SHF backward-wave oscillators the electronic tuning range is just several megahertz per volt of accelerating voltage. Backward-wave oscillator can deliver an output power of tens of

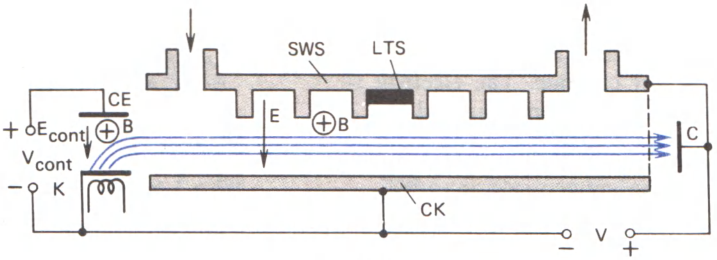


Fig. 20-18
Arrangement of a flat M-TWT

milliwatts to several watts with an efficiency of several per cent. The accelerating voltage usually runs into hundreds or even thousands of volts, and the beam current ranges from units to tens of milliamperes.

There is a modification of backward-wave oscillator tubes, known as *resonant backward-wave tubes*. In them, there is no lossy terminating section, and the slow-wave structure is shorted near the collector so that it can act as a resonant structure. Such devices are amenable to both mechanical and electronic tuning. Resonant backward-wave tubes have a better frequency stability and a higher efficiency.

The magnetrons we discussed earlier deliver an ample power output and are highly efficient, but they have a narrow bandwidth and are not amenable to electronic frequency tuning and gain adjustment. In contrast, O-type TWTs and BWTs have a large bandwidth, lend themselves readily to electronic tuning, and can operate as amplifiers. On the demerit side, they have a low efficiency and will, in some cases, deliver a limited power output. Hence, the development of microwave devices that combine the advantages of magnetrons and TWTs or BWTs.

Wide use is made of M-TWTs and M-BWTs (that is, crossed-field or M-type travelling-wave and backward-wave tubes). Figure 20-18 shows, in sketch form, the arrangement of a flat M-TWT. The electrons emitted by a hot cathode *K* enter the d. c. electric field E_{accel} set up by the voltage existing at the control electrode, *CE*, and a d. c. magnetic field of magnetic flux density *B*, which is at right angles to the plane of the paper and is set up by an external magnetic structure not shown in the figure. These two fields bend the travel path of the electron beam and cause it to move towards the collector *C* through the interaction space between the slow-wave structure *SWS* and a cold cathode *CK*. As is seen, the cold cathode in an M-TWT is located where a hot cathode is placed in magnetrons. The slow-wave structure is held at a constant positive potential with respect to the cold cathode. Therefore, the electron beam is acted upon by the crossed electric field of strength *E* and d. c. magnetic field of magnetic flux density *B*. Travelling in these crossed fields, the electron beam contributes some of its energy to the electromagnetic wave propagated from input to output – that is, amplification takes place. The slow-wave structure includes a lossy terminating section, *LTS*, to prevent the likelihood of spurious oscillations.

With a large input signal, an M-TWT may be 50-70% efficient and have a gain of several hundreds. In CW operation, an M-TWT can deliver several kilowatts of power. In pulse work, the figure may be several megawatts. At this writing, M-TWTs are mainly used as output r. f. power amplifiers. Figure 20-19 shows a cylindrical version of the M-TWT; the notation is the same as in Fig. 20-18.

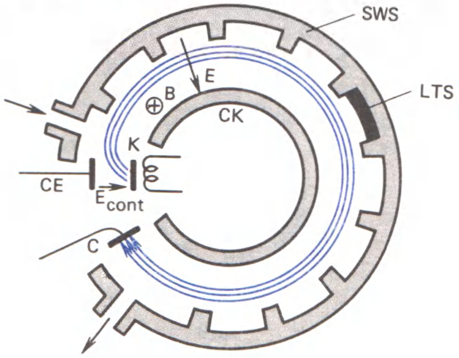


Fig. 20-19
Arrangement of a cylindrical M-TWT

M-BWTs are arranged similarly to M-TWTs, and they may be used as both amplifiers and oscillators. In these tubes, the r.f. output is located close to a hot cathode. The electron beam interacts with the wave propagated towards it. M-BWT amplifiers have both an input and an output while M-BWT oscillators only have an output and also a lossy terminating section near the collector. The CW power output of an M-BWT oscillator is several tens of kilowatts in the UHF band and several hundreds of watts in the SHF band. The efficiency is 50-60%. Electronic tuning can be accomplished by varying the accelerating voltage V.

20-6 The Amplitron and the Carmatron

The *Amplitron* and the *Carmatron* are M-type microwave devices which combine to a certain extent the principles of operation utilized in the magnetron and the M-BWT. In contrast to the M-BWT, however, they have a hot cylindrical cathode similar to that used in the magnetron.

In sketch form, the arrangement of the Amplitron is shown in Fig. 20-20. In contrast to the magnetron, it has a coupled-cavity slow-wave structure which is not closed on itself, that is, it is not re-entrant, and there are an input and an output in the anode block. To avoid stray oscillations in the π-mode (as in the magnetron), the Amplitron usually has an odd number of resonators. As in the magnetron, a closed rotating electron cloud or sheath is formed and interacts with the electromagnetic wave travelling in the opposite direction. The oscillations build up as the beam transfers its energy to the wave.

Amplitrons are mostly used as amplifiers of relatively strong signals in which case they are at least 55-60% efficient. High-power and extra-high-power devices are 70-85% efficient. In CW operation, Amplitrons deliver as much as 500 kW of output power, and in pulse work, 10 MW or even more. Their gain runs into several tens. The relative bandwidth is 5-10%. The beam voltage is units or tens of kilovolts, and the beam current is tens of amperes.

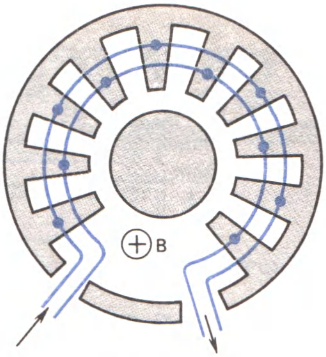


Fig. 20-20

Basic structure of the Amplitron

The Carmatron is a microwave cross-field oscillator which has about the same structure as the Amplitron, except that the input is replaced by a matched load. It has the same power output and efficiency as the Amplitron.

Highly stable oscillations can be generated by an Amplitron ganged up with an external high-Q cavity coupled to the input of the tube, and some extra components. The resultant more sophisticated device is called the *Stabilotron*. It is a self-excited oscillator of an especially high stability, tunable to within 10%.

We have discussed only the most important microwave tubes. There are many more of them, though their use is still limited.

Chapter Twenty-One
Reliability and Testing of Electron Devices

21-1 Reliability and Testing of Semiconductor Devices

As long as they are operated properly, well-manufactured semiconductor devices have a high reliability, their failure rate being $\lambda = 10^{-7}$ or 10^{-8} per hour. The failure rate-vs-time (or 'bath tub') curve shown in Fig. I-1 has a flat portion spanning in the case of semiconductor devices a time interval of tens of thousands of hours. This is far better than tubes are able to show.

Sudden failures in semiconductor devices mainly occur due to the puncture or breakdown of their *p-n* junctions, broken or short-circuited leads. These failures usually account for not more than 20% of the total figure. Gradual failures are far more frequent. In transistors they may be caused by a gradual fall in current gain, an increase in the collector leakage current, and an increase in the noise figure. Sometimes, a temporary instability of device parameters may be observed.

Most often, the reliability of a device is degraded by overheating. Heat build-up is always the No. 1 enemy of semiconductor devices. Therefore, care must always be exercised to keep the case temperature as low as practicable, especially that of high-power devices because they dissipate a good deal of heat. Among other things, devices should, as far as possible, be operated under easy conditions. In many cases, the reliability of semiconductor devices can markedly be improved by using additional heat sinks or radiators. One example is an additional radiator made of metal strips with a hole for a transistor or a diode, as shown in Fig. 21-1. Such a radiator should preferably be made of copper, aluminium, or brass, although steel may also be used. The device case should make a tight contact with the heat sink. Heat withdrawal can be enhanced by increasing the number of fins, and their surface should preferably be blackened.

The reliability of semiconductor devices is adversely affected by moisture. To keep out moisture, devices are enclosed in moisture-tight sealed cases, and the devices themselves are given a thin coat of some protective material.

Although semiconductor devices are mechanically robust and stand up to vibration well, it will be well-advised to protect them against impacts and excessive vibration. As has been noted, semiconductor diodes and bipolar transistors tend to fail when exposed to a strong ionizing radiation. Tunnel diodes and FETs are far more radiation-resistant.

For ample reliability, semiconductor devices should be used subject to a few rules which may be summed up as follows.

1. The operating voltages, currents, and powers should never exceed the respective absolute maximum ratings. A device may never be operated in conditions such that any two parameters reach their absolute maximum ratings at the same time.

2. Transistors may never be operated with their bases open-circuited even for a very short time. For better stability in operation, the resistance of the base circuit should be kept to a minimum.

3. The operating temperature of a device should preferably be kept low. A temperature by 10 degrees Celcius below the limit will halve the number of failures.

4. It will be good practice to safeguard semiconductor devices against overvoltages. This purpose may be served by voltage stabilization circuits. Any supply voltage should never be applied in a wrong polarity.

5. Soldered joints on leads may be made with a soldering iron not more than 60 W in power, with the joint located at least 10 mm away from the device case and completed in a matter of not more than five seconds. When making a soldered joint, it is important to withdraw heat

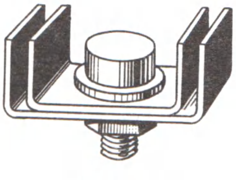


Fig. 21-1

Heat sink (radiator) made of metal plates from between the device case and the joint by, say, clamping the lead in a pair of pliers or forceps.

6. Leads may be bent at a point at least 10 mm away from the device case.

7. Semiconductor devices ought not to be placed near hot or heated components. It will be always a good plan to provide an efficient heat removal from the device case.

8. A device should NEVER be mounted or secured by its leads alone.

9. Semiconductor devices should NEVER be tested with an ohmmeter which can generate currents or voltages dangerous to the device.

10. In the case of MOSFETs, the greatest danger lies in the dielectric layer breaking down, should a charge of static electricity be formed on the gate. Therefore, the gate must always be grounded for d. c. or returned to chassis ground (directly or via a resistor).

Semiconductor devices can best be checked with special-purpose testers. Simple checks, however, have often to be made when no testers are available.

The simplest way to test semiconductor diodes is with an ohmmeter (or a multimeter). The test procedure should include the forward and reverse resistance of the diode. The normal forward resistance is tens of ohms for germanium diodes and hundreds of ohms for silicon diodes. The reverse resistance must be several hundred kilohms for germanium diodes while for silicon diodes it may be several megohms. Higher-power diodes will have lower forward and reverse resistances than lower-power diodes. Instead of the resistances, one may measure the forward and reverse currents with a milliampere and a microampere, respectively. To guard the instruments against damage in the case of a breakdown in the device, a current-limiting series resistor must be connected in the test set-up. Its value can be found by Ohm's law as

$$R_{\text{lim}} = E/I_{\text{max}}$$

where E = supply voltage

I_{max} = maximum current for the instrument.

If the check reveals that the diode resistance is decreasing (the current rising) gradually all the time, the diode should be classed as faulty.

A very simple check for a fault in a *p-n* junction or open-circuits in the leads of a transistor can be run with an ohmmeter. One terminal of an ohmmeter is then connected to the base and the other is touched in turn to the emitter and the collector. The ohmmeter will read either the forward resistance (tens of ohms) or the reverse resistance (hundreds of kilohms or even several megohms), depending on the polarity of the voltage across the junctions. The two resistances of each junction can be determined by interchanging the leads of the ohmmeter. During such a check, NEVER allow the currents and voltages in and across the transistor to exceed their absolute maximum ratings.

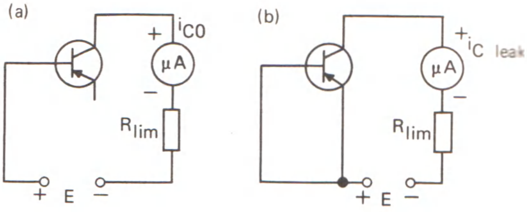


Fig. 21-2

Test set-ups to measure the leakage currents of transistors

The check run as described above does not yet give grounds for saying that the transistor will operate normally in a particular circuit. Therefore, it is important to make sure that the leakage collector current does not exceed the permissible limit and that the beta is not below its normal value.

The leakage currents can be measured, using the test set-ups shown in Fig. 21-2. Both include a current-limiting resistor as a way of preventing any damage to the microammeter. The value of this resistor can be found as explained earlier. If all leakage currents can be measured, it will suffice to measure I_{C0}. Where high-power transistors are involved, a milliammeter should be used instead of a microammeter. When measuring the leakage currents, a gradual rise in the respective current is an indication that the transistor is at fault.

The beta gain can in the simplest case be measured with the test set-up shown in Fig. 21-3. Here the supply voltage E is several volts, and the base-lead resistor R_B is such that the base current will have some normal value, i_B, specified for the transistor in question. It is legitimate to deem that $R_B \approx E/i_B$. Dividing i_C as measured with the milliammeter by the base current will yield an approximate value of the beta:

$\beta \approx i_C/i_B$

Or, after replacing i_B with E/R_B, we will have

$\beta \approx i_C R_B/E$

For example, if $E = 4.5$ V and $i_B = 0.1$ mA, then

$R_B \approx 4.5 \div 0.1 = 45$ kΩ

At $i_C = 4$ mA, we will have

$\beta = 4 \div 0.1 = 40$

The error with this technique is small because it does not take into account the voltage drop across the emitter junction which is only a few tenths of a volt, that is, a small fraction of *E*. No high accuracy is generally required in measuring the beta gain because transistors always show a marked spread in parameters between units.

It would be a good idea to test a transistor in some simple oscillator circuit, but this test is not mandatory.

FETs should be checked for the channel conductance between the source and the drain and also for proper functioning of the gate junction. The latter should be tested as the *p-n* junctions of diodes or bipolar transistors. Of course the test must be run at low voltages so as not to cause the *p-n* junction to break down.

In MOSFETs, it is important to check the channel conductance and the insulation between the gate and the channel. In all forms of testing, the channel must be grounded, lest the dielectric layer should break down. The quality of a FET can be ascertained by measuring its transconductance which is the key parameter of these devices. The transconductance can be determined by measuring the drain current while varying the gate voltage. Notably, the voltage can be varied by placing a 1.5-V dry cell between the gate and the source. Variations in the drain current can then be measured with a milliammeter.

21-2 Reliability and Testing of Tubes

Sudden failures in tubes may be caused by a short-circuit between the electrodes, by a broken wire, by the rupture of the insulation, by cracks in the tube envelope, and some other events. Gradual failures result from slow irreversible changes in the oxide cathode leading to a fall in emission, interelectrode leakage, release of gas by the electrodes, etc.

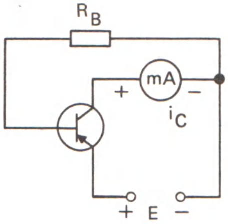


Fig. 21-3
Simple test set-up to measure the beta current gain of transistors

The failure rate for tubes is typically 10^{-5} per hour, or even less. Tubes with normal and enhanced reliability and service life differ in failure rate by a factor of 5 to 10 or even more. Least reliable of all are high-power transmitting, modulator and amplifying tubes, H. V. rectifiers (kenotrons), and other high-power tubes. High reliability and long life can be assured on abiding by all the service rules set forth in makers' manuals. Above all, one ought not to exceed the applicable absolute maximum ratings for current, voltage and power, and also the limits for ambient factors, such as temperature, pressure and relative humidity, the intensity of impacts and vibration, and other mechanical influences. Tubes may never be operated in conditions when any two parameters may reach their absolute maximum ratings at the same time.

Overheating is a major cause of failures. Obviously, for a tube to be as reliable as practicable, its heating must be kept to a reasonable minimum. A reduction of just a few degrees in a tube's temperature may prove decisive for its reliability. It will always be good practice to run a tube under conditions which reduce its heating. Good heat abstraction from the tube is likewise desirable. Sometimes it will prove advantageous to put a finned radiator over the envelope of a tube that is heavily heated in service. The radiator may be fabricated from a metal such as aluminium, brass or copper (Fig. 21-4). The outer surface of the radiator should preferably be blackened so as to enhance heat abstraction by radiation. Of course, measures should likewise be taken to minimize heat input from other components, sun light, and any other external sources. It is important to remember that large doses of ionizing radiation may degrade a tube's performance. The fungi and high relative humidity most likely to exist in the tropics will also impair the reliability of contact in tube sockets.

In service, the following factors may degrade the reliability of tubes:

- the filament (or heater) voltage is a maximum while the cathode current is small;
- the filament (or heater) voltage is a minimum while the cathode current is large;
- the electrode dissipation is a maximum while the control grid circuit presents a high resistance;
- the envelope temperature is a maximum and the electrode voltages are high while the cathode current is low.

Care must be taken to minimize vibration, impacts and other mechanical influences. When a tube is operated at a reduced ambient pressure, heat abstraction is impaired, and steps must be taken to bring down the electrode dissipation. An elevated relative humidity may cause oxidation and poor contact in the tube sockets, increased leakage current, and even an inter-lead breakdown.

It is important to insert and secure tubes properly. Data sheets often specify a vertical position for a given tube as the only one allowable, and this advice must be followed. When making soldered joints on the leads of miniature tubes, a good heat removal must be provided between the joint and the envelope. This can be done by, say, clamping the lead in a pair of pliers. Leads may be bent at least 5 mm away from the envelope.

If tubes are to operate reliably and for a long time, it is essentially important to abide by the rules and regulations both set forth above and given in service or operation manuals.

When a malfunction is detected in a tube-based electronic or telecommunication equipment, trouble-shooting must be started by checking the tubes because they are the most common cause of many failures. Most receiving-amplifying tubes can be checked with tube testers. How to use these testers is usually explained in the tutorial material included in the maker's manual. If a tube tester is not available, resort may be made to the simpler techniques of tube testing. One such technique is to insert the tube being tested in a similar circuit known to operate trouble-free. Then it will be immediately seen whether the tube is at fault or not.

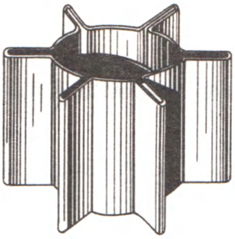


Fig. 21-4

Sheet-metal radiator to cool a tube

It is useful to learn how to check tubes without resort to any circuit. For example, a multimeter can be employed to see if the heater or directly-heated cathode is intact, or whether there is a short between the electrodes of a tube. As an alternative, this can be done with a make-shift tester consisting of a series combination of a current source (say, a dry cell) and a voltmeter. Instead of a voltmeter, use may be made of a milliammeter and a series resistor, or a pair of headphones, or an incandescent lamp.

Cathode emission can be checked, using the test set-up shown in Fig. 21-5. The procedure is as follows. Apply the normal heater voltage, connect all the grids to the anode and operate them as a single anode, making sure that the anode supply voltage is not more than 10-15 volts. If there is emission from the cathode, the milliammeter connected in the anode circuit will read a current. The milliammeter may be replaced with a voltmeter. When this technique is used to determine the emission of a tube known to be good, the indication of the instrument will give a measure of the emission in any other tubes of the same type. This test may well be run without an anode supply, if the anode circuit is connected to the "+" terminal of the heater battery, but the anode current will then be smaller by an appreciable amount.

The leads of a tube can be checked for breaks ('open-circuits') by placing the milliammeter in the set-up of Fig. 21-5 in series with each electrode in turn (in the diagram, the points of connection are labelled with crosses). If there is no break (or 'open-circuit') in a given lead, the instrument will indicate that there is a current flowing in the wire of the respective electrode.

Since transconductance is the principal parameter of amplifying tubes, the respective check is desirable. If the check shows that the transconductance has a normal value, this is an indication that the tube is healthy. The transconductance of a tube is checked by feeding normal supply voltages to the tube's electrodes (making sure that the respective currents and powers do not exceed their absolute maximum ratings). A milliammeter must be placed in the anode circuit. After the control-grid voltage has been changed by 1 or 1.5 volts (by, say, connecting one dry cell in the control grid lead), note the change in the anode current. From these figures, it is a simple matter to determine the transconductance of the tube. Another method consists in that a low-value load resistor (say, 100 Ω) is connected in the anode circuit, and the control grid is fed with an alternating sinewave voltage of a known value. Then the amplified voltage is measured across the load resistor. Dividing it by the load resistance will yield the value of the alternating anode current. Now it is easy to determine the transconductance. Because of the load, the transconductance thus determined will be somewhat smaller than the true, or static, transconductance of the tube.

With tube testers, all the checks are carried out in a similar manner, but with a greater convenience to the operator because all the necessary changes in the circuit are made by manipulating appropriate switches.

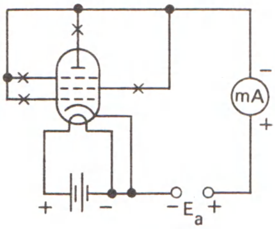


Fig. 21-5

Simple set-up to test tube leads and cathode emission

Glow-discharge tubes (such as neon bulbs, gas-filled rectifier diodes, thyratrons, character-display tubes and the like) should be tested for the starting (or firing) voltage and the presence of glow. When carrying out a test, it is important to use a current-limiting resistor, lest the glow discharge should change into an arc discharge. Because the current drawn by glow-discharge tubes is ordinarily several milliamperes, the resistance of the current-limiting resistor can always be found by Ohm's law: take a current of 2 or 3 mA and divide the difference between the supply voltage and the operating voltage of the tube in question by this current.

Part Three
Photoelectric and Optoelectronic Devices

Chapter Twenty-Two
Photoelectric Devices

22-1 Photoelectric Emission

Photoelectric emission, also termed *photoemission* or the *outer photoemissive effect*, refers to the liberation of electrons from matter when it is exposed to electromagnetic radiation of certain energies. The emitting electrode of a photoelectric device is then called a *photoemissive cathode* or, simply, a *photocathode*, and the electrons thus emitted are named *photoemissive electrons* or *photoelectrons*.

Studies into photoemission go back to 1886 when Heinrich Hertz noticed that a lower voltage was required to set off an electric discharge between a pair of electrodes when one of the two electrodes was illuminated. In 1888, A. G. Stoletov of Moscow University (Russia) began a systematic study into the effect. He described most of the important properties of the outer photoemissive effect but was unable to explain it because electrons had not yet been discovered.

What follows is a brief summary of the laws and relations bearing on photoemission.

1. *The Stoletov law.* The photocurrent I_{ph} arising due to photoemission is proportional to the incident luminous flux Φ:

$$I_{ph} = S\Phi \qquad (22\text{-}1)$$

where S = sensitivity of the photocathode, usually expressed in microamperes per lumen (the lumen is the unit of luminous flux).

If the incident luminous flux Φ is *monochromatic*, that is, contains electromagnetic radiation of only one wavelength (or, which is the same, of only one frequency), we have what is known as the *monochromatic sensitivity* symbolized as S_λ. The sensitivity towards a flux of white (nonmonochromatic) light consisting of radiation of all likely wavelengths (frequencies) is referred to as the *integral sensitivity* symbolized as S_Σ.

2. *The Einstein equation.* In 1905, Albert Einstein suggested that radiant energy was transmitted by photons, each with an energy $h\nu$ and localized in space so that each is capable of reacting with an electron. The photons are capable of transferring all of their energy to electrons after which they disappear. Of the energy $h\nu$ transferred from the photon to an electron, a certain amount W_0 may be used up in overcoming the potential barrier at the surface of the material. This amount W_0 is the *photoelectric work function* of the material, $W_0 = h\nu$. The maximum kinetic energy that any electron may have after photoemission is then

$$E = h\nu - W_0 \qquad (22\text{-}2a)$$

which is known as the *Einstein photoelectric equation.* Here,

$$h\nu = W_0 + 0.5mu^2 \qquad (22\text{-}2b)$$

where m = mass of a photoelectron
u = velocity of a photoelectron
ν = frequency of the incident radiation
h = Planck's constant equal to $6.63 \times$ $\times 10^{-34}$ J s

As will be recalled, there are two aspects of electromagnetic radiation used together or separately to explain many phenomena associated with light. On the one hand, it consists of waves each of wavelength λ and of frequency ν. On the other, it may be treated as a flux of discrete quanta or packets, called *photons*, each of energy $h\nu$.

3. In the case of photoemission from solids, electrons are only liberated when the frequency of the incident radiation is greater than a characteristic value – the *photoelectric threshold frequency* of the material, designated as ν_0. Below that frequency, photoemission ceases because $h\nu_0 = W_0$ (where W_0 is the *threshold photoelectric work function*) and the kinetic energy of photoelectrons becomes equal to zero. The wavelength corresponding to ν_0 is $\lambda_0 =$ $= c/\nu_0$, where c is the velocity of light equal to 3×10^8 m s^{-1}. At a frequency below the photo-

electric threshold, that is, when $\nu < \nu_0$ or $\lambda > \lambda_0$, photoemission cannot take place because $h\nu < h\nu_0$ which tells us that the energy of the photon is not enough for an electron to overcome the potential barrier at the surface of the material.

4. Photoemission shows a very negligible time lag – variations in the photocurrent only slightly lag behind variations in the incident radiation (by as little as a few nanoseconds).

Apart from monochromatic and integral sensitivities, photocathodes are sometimes characterized in terms of the ratio between the number of emitted photoelectrons and the number of photons causing the photoemission. This ratio has come to be called the *electron quantum yield* or *quantum efficiency*.

If each of all available photons caused the liberation of one electron, the quantum yield would be equal to unity. However, the greater proportion of photons do not contribute to photoemission: some of them have a wavelength in excess of λ_0, others move very deep into the cathode and dissipate their energy there, and still others are reflected from the cathode surface. As a rule, the quantum yield or efficiency does not exceed 1-2%.

The photoelectric work function W_0 and the photoelectric threshold wavelength λ_0 for some elements are given below.

	Cesium	Potassium	Antimony	Germanium	Silicon
W_0, eV	1.9	2.3	4.0	4.4	4.8
λ_0, μm	0.66	0.55	0.31	0.28	0.21

Visible light falls within the spectrum region from 0.38 to 0.78 μm, so, as follows from the above table, it can only cause photoemission from cesium and potassium. Therefore, photocathodes are usually fabricated from other than pure metals. For example, one of the most commonly used photocathode types consists of a mixture of silver, cesium oxide, and pure cesium. This reduces its work function, and its photoelectric threshold wavelength is $\lambda_0 = 1.1$ μm.

The sensitivity of a photocathode is a function of the wavelength of the incident radiation. This relationship, $S = f(\lambda)$, is called the *spectral response* of a photocathode and may take any one of two forms as shown in Fig. 22-1. Curve *1* applies to the 'normal' photoelectric effect which occurs with substantial cathodes fabricated from pure metals, and curve *2* applies to what is known as the selective photoelectric effect typical of thin cathodes fabricated from specially treated alkali metals. With time, the sensitivity of photocathodes diminishes – a condition which is referred to as *photocathode fatigue*.

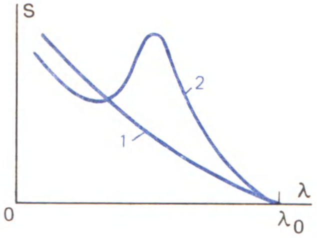


Fig. 22-1
Spectral responses of photocathodes

22-2 Phototubes

A *phototube*, which may be the *vacuum type* or the *gas-filled type*, is a diode with the inner surface of its envelope given a thin coat of a material capable of emitting photoelectrons. This coating is the photocathode of the diode. Its anode is usually a metal ring which does not stand in the way of light incident on the photocathode. Vacuum phototubes, as their name implies, use a high vacuum inside their envelopes. Gas-filled (or simply gas) phototubes have an inert gas, such as argon, which fills the envelope at a pressure of several hundred pascals (a few mm Hg). The photocathode is ordinarily made of antimony-cesium or silver-oxygen-cesium.

The performance of phototubes is stated in terms of their characteristics. Figure 22-2*a* shows the *anode* (*current-voltage*) *characteristics* of a vacuum photodiode, $I_{ph} = f(v_a)$ at $\Phi = $ const. They show a well-defined saturation condition. The same characteristics of a gas phototube first run in the same manner as for vacuum phototubes, then rise at an ever increasing rate because the further increase in anode voltage leads to the ionization of the filling gas, and the current builds up appreciably (Fig. 22-2*b*). This situation is stated in terms of the *gas amplification* (or *gas multiplication*) *factor* which may range from 5 to 12. Figure 22-3 shows the current-vs-luminous flux characteristics, $I_{ph} = f(\Phi)$ at $V_a = $ const. There are also *frequency response curves* which relate the sen-

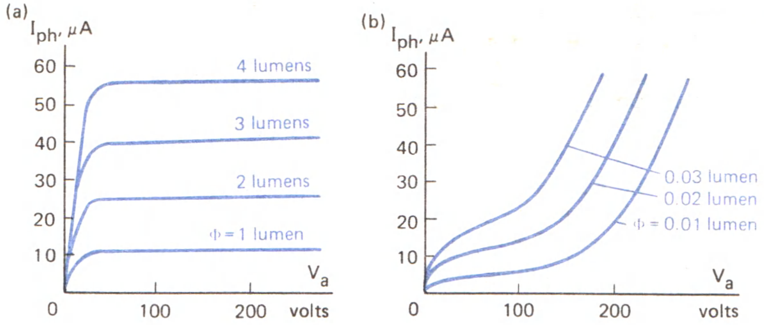


Fig. 22-2

Anode characteristics of (*a*) a vacuum phototube and (*b*) a gas-filled phototube

sitivity of a photocathode to the frequency of the incident luminous flux. It is seen from Fig. 22-4 that vacuum phototubes (line *1*) possess a small time lag. They are able to operate at frequencies of several hundred megahertz. Gas phototubes (curve *2*) show an appreciable time lag and their sensitivity falls off at frequencies as low as a few kilohertz.

The principal parameters of phototubes are cathode sensitivity, maximum allowable anode current, and dark current. The cathode sensitivity of vacuum phototubes is tens of microamperes per lumen. In the case of gas phototubes the figure is several hundred microamperes per lumen. The dark current is the d. c. current that flows without any light. It owes its origin to the thermionic emission of the cathode and the leakage currents between the electrodes. At room temperature, the thermoemission current may be 10^{-10}-10^{-11} A, and the leakage currents may be 10^{-7}-10^{-8} A. The leakage currents can be minimized by giving a tube a special design, while the thermoemission current can be brought down only by cooling the photocathode to a very low temperature. The dark current limits the use of phototubes as detectors of very weak light signals.

It is usual to connect a phototube in series with a load resistor, R_L (Fig. 22-5). Since the photocurrent is very small, the d. c. resistance of a phototube, R_0, is very high and amounts to units or even tens of megohms. Therefore, the load resistance should likewise be very high. The voltage generated by the light signal is picked off the load resistor and is fed to an amplifier whose input capacitance shunts R_L. The higher the value of R_L and the higher the signal frequency, the stronger is this shunting action and the lower the useful signal voltage across R_L.

Phototubes have found many uses in automatic process control, sound motion pictures, and physical research instrumentation. Unfortunately, they cannot be microminiaturized and need relatively high anode voltages (tens or even hundreds of volts). Because of this, they have been ousted from many applications by semiconductor light detectors some of which will be described in the next chapter of this book.

22-3 The Photomultiplier Tube

A *photomultiplier tube* is essentially a vacuum phototube supplemented by a device to amplify the photocurrent due to secondary electron emission.

In sketch form, the arrangement of a photomultiplier tube is shown in Fig. 22-6. The vacuum envelope contains a photocathode, *PC*, and a series of anodes, called *dynodes*, D_1

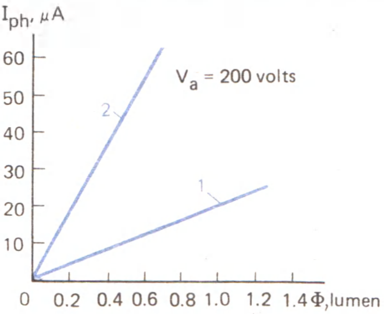


Fig. 22-3

Current-vs-flux characteristics of a vacuum phototube (*1*) and a gas-filled phototube (*2*) through D_3, each at a progressively higher potential. The incident luminous flux Φ brings about photoemission from the photocathode, and an accelerating electrostatic field urges the emitted photoelectrons towards the first dynode which is positive with respect to the photocathode. The dynode is made of a material capable of a sufficiently strong and stable secondary electron emission. The primary electrons that constitute a photocurrent I_{ph} knock out of the dynode secondary electrons whose number is σ times the number of primary electrons (σ is the secondary-emission ratio, usually equal to several units for the first dynode). Therefore, the current constituted by the secondary electrons emitted by the first dynode is $I_1 = \sigma I_{ph}$. The secondary electrons created at the first dynode are attracted to the second dynode which is more positive than the first, and again knock out of it secondary electrons which constitute a current I_2 which is σ times I_1 (for simplicity, it is assumed that the secondary emission ratio is the same for all the dynodes), that is, $I_2 = \sigma I_1 = \sigma^2 I_{ph}$. The secondary electrons created at the second dynode are attracted to the third dynode still more positive than the second, and the electrons created there leave as a current $I_3 = \sigma I_2 = \sigma^3 I_{ph}$, and so on. The secondary electrons created at the last, nth, dynode, D_n, and constituting an electron current I_n, are attracted to the anode, A, and the anode current is then

$$I_a = I_n = \sigma^n I_{ph}$$

Thus, theoretically, the current gain is $k_I = \sigma^n$. For example, if $\sigma = 10$ and $n = 8$, then k_I will be 10^8. Practically, the total current multiplication is smaller because not all the secondary electrons knocked out of one dynode are collected by the next dynode. To maximize the number of secondary electrons collected by the succeeding dynodes and finally by the anode, several designs of photomultiplier tube have been developed in which the electrodes are given a suitable shape and appropriately disposed relative to one another. The stream of secondary electrons travelling from one dynode to the next is usually focused by an electrostatic field because magnetic focusing would involve the use of a bulky magnetic system.

The simplest photomultiplier tube is one having a photocathode, a single dynode, and an anode. This is a single-stage photomultiplier tube. Multistage photomultiplier tubes can give a total current multiplication of several millions, and their integral sensitivity may be as high as tens of amperes per lumen. As a rule, photomultiplier tubes operate at an anode current of not more than several tens of milliamperes, and the input luminous flux may be small as 10^{-5}-10^{-3} lumen or even smaller.

Since the consecutive dynodes are each at a progressively higher potential, the anode voltage has to be as high as 1 or 2 kV, which is a limitation of photomultiplier tubes. As a rule, a phototube is powered from a voltage divider energized with the total anode voltage (Fig. 22-7). The anode circuit contains a load resistor, R_L, from which the output voltage is picked off.

As is the case with conventional phototubes,

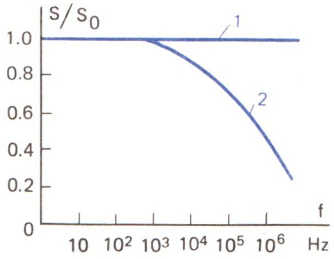


Fig. 22-4

Sensitivity-vs-frequency response of (*1*) a vacuum phototube and (*2*) a gas-filled phototube

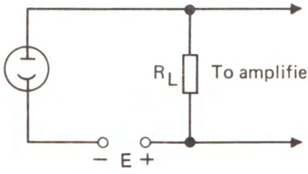


Fig. 22-5

Connection of a phototube in a circuit

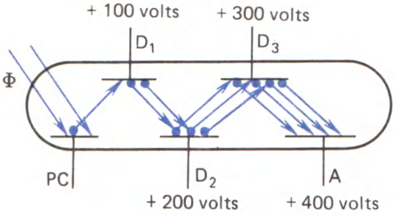


Fig. 22-6

Arrangement and operation of a photomultiplier tube the sensitivity of photomultiplier tubes is limited by dark current which is mainly due to the thermionic emission from the photocathode and the dynodes. Its value is not more than a fraction of a microampere. This current can be minimized by cooling the photomultiplier tube.

As already noted, dark current limits the minimum change in the incident luminous flux that can be detected by a photomultiplier tube. The valid changes in the luminous flux are in turn limited by fluctuations in the photocathode emission and dark current. It should be added, however, that these fluctuations are small, so photomultiplier tubes are low-noise devices. Their noise figure is not more than 1.5 or 2 (as will be recalled, the noise figure of an ideal noise-free amplifier is unity).

The key parameters of photomultiplier tubes are the spectral region (the wavelength range) within which a given device can be used to advantage, the number of multiplication stages, the total current multiplication, the supply voltage, the integral sensitivity, and the dark current. It is usual to describe the performance of a photomultiplier tube by reference to its light response curve, $I_a = f(\Phi)$, and also plots relating the current multiplication k_I and the integral sensitivity S_Σ to the anode supply voltage E_a (Fig. 22-8).

Photomultiplier tubes show a negligible time lag and may be used at fairly high frequencies. They can be used to detect light pulses recurring at nanosecond intervals.

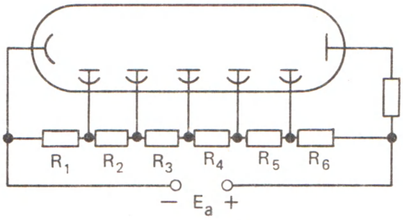


Fig. 22-7

Connection of a photomultiplier tube in a circuit

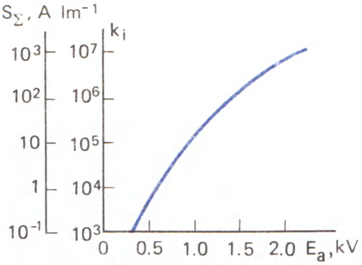


Fig. 22-8

Current gain and integral sensitivity as functions of the supply voltage of a photomultiplier tube

There is no semiconductor device that could do the job of the vacuum photomultiplier tube. Therefore, it is still being used in many fields of science and technology, including astronomy, facsimile telegraph, television, measurement of small luminous fluxes, spectral analysis, etc.

Chapter Twenty-Three

Semiconductor Optoelectronic Devices

23-1 General

In the last few years optics and electronics have merged to produce a new technology generally called *optoelectronics* or, sometimes, *electro-optics*. It encompasses a wide range of semiconductor light detectors (photoresistors, photodiodes, phototransistors, and photothyristors) which depend for their operation on what is called the *inner photoelectric effect*.

The inner photoelectric effect, also called the *photoconductive effect*, refers to the increase in conductivity observed in certain elements and compounds exposed to electromagnetic radiation. More specifically, this occurs due to the generation of electron-hole pairs in semiconductors. The additional conductivity resulting from the action of photons is called *photoconductivity*. Metals do not practically show photoconductivity because their concentration of conduction electrons is huge indeed (about

10^{22} cm^{-3}) and it cannot rise markedly under the influence of the incident radiation. In some devices, the photogeneration of electron-hole pairs gives rise to what is known as the *photo-emf*. This is the *photovoltaic effect*, and photovoltaic devices operate as current sources. On recombining, the electrons and holes in a semiconductor produce photons, so given certain conditions, such semiconductor devices are able to operate as light emitters.

In the sections that follow, we will dwell in more detail on the most commonly used semiconductor devices operating as radiation (or light) sources or photodetectors, or their combinations referred to as *optoisolators* or *optocouplers*, depending on how one looks at what they are intended to do.

Many of the devices described in this chapter are commercially available in both discrete form and as elements of integrated circuits.

23-2 Bulk Photoconductors (Photoresistors)

A *bulk photoconductor, or a photoresistor*, is a semiconductor device whose conductivity increases (whose resistivity decreases) in proportion to the intensity of incident light.

In sketch form, the arrangement of a bulk photoconductor (or photoresistor) is shown in Fig. 23-1*a*. Notice that there is no *p-n* junction necessary for the operation of these devices, just a layer of photoconductive semiconductor material *2* is deposited on a dielectric substrate *1*. Metallic electrodes *3* and a suitable enclosure with a window are then added to complete the device. The connection of a bulk photoconductor in a circuit is shown in Fig. 23-1*b*. The polarity of the power supply is of no importance.

As long as no light is incident upon the device, the bulk photoconductor has a certain high resistance called the *dark resistance*, R_d. It is one of the parameters of photoresistors and may be as large as 10^4-$10^7 \Omega$. The associated current flowing through the device is called the *dark current*. When there is incident radiation with photons having a sufficient energy, pairs of mobile carriers (electrons and holes) are generated in it, and its resistance decreases.

Several semiconductor materials having some desired properties are used to manufacture photoresistors. For example, lead sulphide is most sensitive to infra-red light, and cadmium sulphide to visible light. The performance of a photoresistor is usually assessed in terms of the *specific sensitivity*, that is, the integral sensitivity per volt of applied voltage:

$$S_{sp} = I/\Phi V \qquad (23\text{-}1)$$

It is usually several hundred or thousand microamperes per volt per lumen.

Photoresistors have a linear *V-I* characteristic and a nonlinear current-flux characteristic (Fig. 23-2).

In addition to the dark resistance and specific sensitivity, the parameters of a photoresistor include the maximum allowable operating voltage (usually 600 V), the ratio of dark to light resistance (which may be as great as 1000 : 1 or even greater), and the temperature coefficient of photocurrent, TCPC = $\Delta I/I\Delta T$. A limitation of photoresistors, similarly to that of all semiconductors, is the heavy dependence of resistance on temperature. Another limitation is a noticeable time lag – after the incident radiation has been turned off, the electrons and holes will

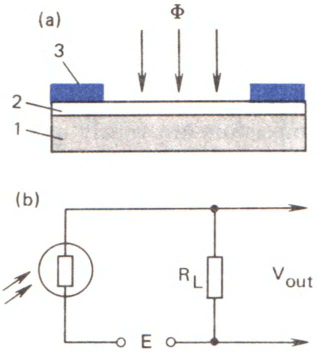


Fig. 23-1

Structure and connection of a photoresistor (bulk photoconductor) in a circuit

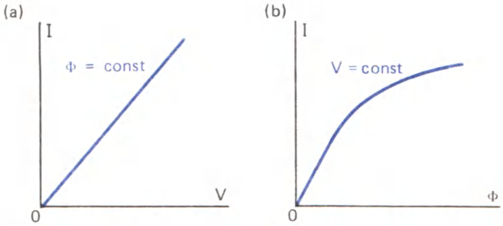


Fig. 23-2

Photoresistor: (*a*) current-voltage characteristic and (*b*) current-flux characteristic go on recombining for some time. Practically, photoresistors are used at frequencies not higher than several hundred hertz or several kilohertz. Inherent noise in photoresistors is considerable. Yet, photoresistors are widely used in various automatic control circuits and other applications.

23-3 Photodiodes

A *photodiode* is a semiconductor diode which depends for its operation on the inner photoelectric effect. The incident light flux controls the reverse current of a photodiode. When photons of energy greater than the energy gap of the device material are absorbed in the device, hole-electron pairs are generated, the conductivity of the diode increases, and the reverse current builds up. This is the *current mode* of operation (Fig. 23-3). The current-voltage characteristics, $I = f(V)$, at Φ = const for the current mode of operation (Fig. 23-4) are not unlike the output characteristics of a bipolar transistor connected in a common-base circuit. When there is no light flux incident on the device, the usual reverse leakage current, called the *dark current*, is flowing through the device. When, however, there is a light flux incident on the device, the current through the photodiode increases, and the characteristics run higher. The stronger the incident radiation, the greater the current. An increase in the reverse voltage across a photodiode will raise the current only slightly. At some applied voltage, however, there occurs a breakdown (the dashed portions of the characteristics). The current-flux characteristics, $I = f(\Phi)$, of photodiodes at V = const are linear and almost independent of the applied voltage (Fig. 23-5).

The integral sensitivity of photodiodes is usually tens of milliamperes per lumen. It depends on the wavelength of the incident radiation and has a peak at some certain wavelength which differs from one semiconductor material to another. Photodiodes show a negligible time lag, so they may be operated at frequencies of several hundred megahertz. The frequency limit for *p-i-n* photodiodes is as high as tens of gigahertz. The operating voltage of photodiodes is usually 10 to 30 volts. The dark current does not exceed 10-20 μA for germanium devices and 1-2 μA for silicon devices. The light current is several hundred microamperes. Recently, new types of photodiodes have been developed, using composite semiconductor materials most sensitive to infrared light. Most photodiodes are manufactured by planar technology (Fig. 23-6).

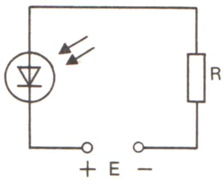


Fig. 23-3

Connection of a photodiode for operation in the current mode

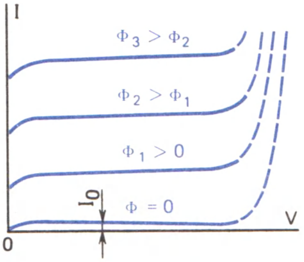


Fig. 23-4

Current-voltage characteristics of a photodiode in the current mode

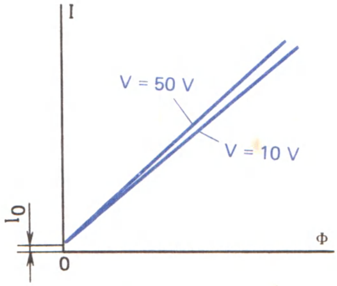


Fig. 23-5

Current-vs-flux characteristics of a photodiode

Photodiodes generally come in several basic types. In *avalanche photodiodes*, the carriers are generated at the *p-n* junction by the avalanche mechanism which substantially raises the device's sensitivity. In *Schottky barrier photodiodes*, the semiconductor is brought in contact with a metal owing to which they show a high speed of response. Improved performance is displayed by *heterojunction photodiodes*. All photodiodes are able to operate as emf or voltage generators. This matter is discussed in the next section.

23-4 Semiconductor Photovoltaic Cells

Semiconductor *photovoltaic cells* serve to convert radiant energy into electric energy. In fact, they are photodiodes operating with no external applied bias and generating an emf of its own due to the effect of incident radiation. This is the *voltage mode of operation* for photodiodes.

Photons incident on the *p-n* junction and the adjacent regions give rise to the generation of hole-electron pairs. The electrons and holes created in the *n*- and *p*-regions diffuse towards the junction and, if they have no time to recombine in transit, they fall under the influence of the internal electric field existing in the junction. This field also acts on the carriers created in the junction itself. The field separates the electrons from the holes. With regard to the majority carriers, however, (for example, the holes in the *p*-region) the field is an accelerating one. It sweeps the electrons into the *n*-region. Similarly, the holes are swept by the field out of the *n*- into the *p*-region. With regard to the majority carriers, however, (for example, the holes in the *p*-region) the field of the junction is a decelerating one, and these carriers stay in the respective region, that is, the holes remain in the *p*-region and the electrons in the *n*-region (Fig. 23-7).

As a result, excess majority carriers accumulate in their respective regions. That is, the electrons and holes build up charges in the respective regions, and there appears a difference of potential called the *photo-emf*, E_{ph}. This photo-emf builds up nonlinearly with increasing luminous flux, or *irradiance* (Fig. 23-8). This emf may be as great as several tenths of a volt. When a semiconductor photovoltaic cell is connected to a load (Fig. 23-9), a photocurrent, I_{ph}, begins to flow, equal to $E_{ph}/(R_L + R_i)$, where R_i is the internal resistance of the photovoltaic cell itself.

The early photovoltaic cells were made of copper hemioxide in 1926. They were followed by photovoltaic cells made of *p*-type selenium. Actually, a thin *n*-region was made in *p*-type selenium to form a *p-n* junction, and the *n*-region was exposed to incident radiation. The integral sensitivity of selenium-junction photocells was as large as several hundred microamperes per lumen. They had a spectral response very close to that of the human eye – a feature of special value for photometric studies. Special mention should be made of thallium-sulphide junction photocells – they had a sensitivity of several thousand microamperes per lumen. Among the disadvantages of photovoltaic cells is a poor frequency response and a relatively strong dependence of the integral sensitivity on temperature.

At this writing, this field is led by silicon-junction photovoltaic cells employed as devices for the direct conversion of solar energy to electric energy. Quite aptly, they are called solar cells. The photovoltaic solar cell has become the workhorse for producing spacecraft power, and solar-cell technology has found its way into

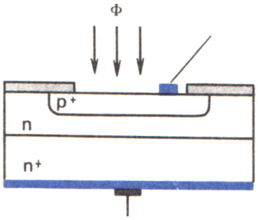


Fig. 23-6
Structure of a planar photodiode

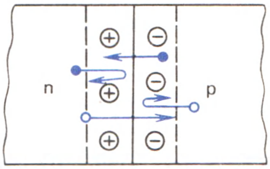


Fig. 23-7
Separation of light-induced carriers by the field of a *p-n* junction

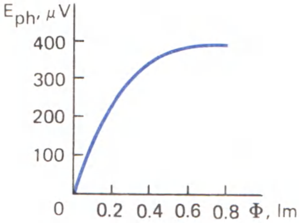


Fig. 23-8
Photo-emf as a function of light flux

many Earth-based applications, such as the heating of houses. Solar cells are able to generate an emf of 0.4-0.5 volt per cell. The individual cells can be connected in parallel and series into solar batteries or panels which have a relatively high efficiency (up to 20%) and can deliver a power output of several kilowatts.

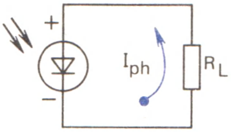


Fig. 23-9

Connection of a photovoltaic cell

23-5 Phototransistors

A *bipolar phototransistor* is similar in construction to a conventional bipolar transistor, except that the top surface has a window or lens so that its base region can be exposed to incident radiation.

Photons incident on the base bring about the generation of electron-hole pairs which diffuse towards the collector junction where they are separated in much the same way as they are in a photodiode. The collector field urges the holes from the base to the collector where they augment the base current. The electrons, in contrast, remain in the base and raise the forward voltage across the emitter junction – a factor which stimulates the injection of holes in this junction, thereby building up the collector current.

The chain of events we have just described applies when the phototransistor is the *p-n-p* type with its base 'floating', that is, not connected to any circuit, as shown in Fig. 23-10, which makes it a two-lead device. As usual, the emitter junction is forward-biased and the collector junction is reverse-biased.

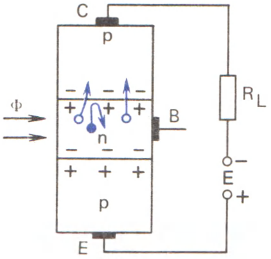


Fig. 23-10

Structure and connection of a floating-base phototransistor

The integral sensitivity of phototransistors may be tens of times that of photodiodes and may run into several hundred milliamperes per lumen.

Floating-base phototransistors are subject to the same temperature variations as conventional bipolar transistors. This drawback is usually counteracted through the use of temperature-compensation schemes some of which have been examined in Chap. 4. Of course, in such cases, the base lead has to be used and so we have a three-lead device. This lead may also be used in order to apply a constant bias voltage or electric signals so that they can act on the device jointly with light signals.

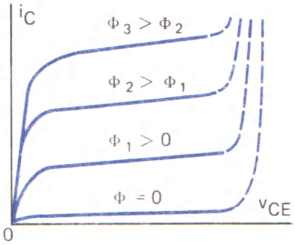


Fig. 23-11

Output characteristics of a phototransistor

The output characteristics of a phototransistor are shown in Fig. 23-11. They are similar to those of a conventional transistor connected in a common-emitter circuit, but each curve corresponds to a particular level of irradiance (luminous flux) and not to a particular value of base current. These characteristics tell us that when the collector-to-emitter voltage, v_{CE}, exceeds a certain value, there occurs an electric breakdown (the dashed portions of the curves).

The parameters of phototransistors are the integral sensitivity, the operating voltage (which usually is 10-15 V), the dark current (which may run into several hundred microamperes), the operating current (up to tens of microamperes), the maximum allowable power dissipation (up to tens of milliwatts), and the upper frequency limit. Phototransistors fabricated by the alloying method have an upper frequency limit of several kilohertz, and those manufactured by the diffusion method (planar phototransistors) are able to operate at frequencies up to several megahertz. A limitation of phototransistors is a relatively high inherent noise level.

The bipolar design is not the only member of the phototransistor family. Another member is the *compound phototransistor*. In effect, this device is a combination of a phototransistor and a conventional bipolar transistor. It has been shown in Chap. 9 that a compound-connected transistor has a current gain, β, equal to the product of the two transistors' betas, that is, $\beta_1 \beta_2$. As a result, the integral sensitivity of a compound-connected phototransistor is tens of times the figure assured by a conventional device and thousands of times the figure obtained with photodiodes. High sensitivity and high speed are achieved by a combination of a photodiode and r. f. transistor.

In addition to bipolar phototransistors, there are also *photofield-effect transistors* or *photo-FETs* which are similar to conventional JFETs, with the exception that they have a lens for focusing light onto the gate junction. As an example, Fig. 23-12 shows the arrangement of an *n*-channel photoFET. When the *n*-channel is illuminated, electrons and holes are produced in the channel and the adjacent *p*-region (the gate junction). The junction between the *n*-channel and the *p*-region is reverse-biased, and so the field of this junction separates the electrons from the holes. This leads to an increase in the electron concentration in the *n*-channel, a reduction in its resistance, and an increase in the hole concentration in the *p*-region. The channel (or drain) current rises. Also, a photon-induced current begins to flow in the gate circuit. This current produces a voltage drop across R_G owing to which the reverse bias voltage across the channel-gate junction is reduced. This leads in turn to an increase in the channel width and, in consequence, to a further decrease in its resistance and a further rise in the drain current. In this way, incident light controls the drain current.

Of special interest are *depletion-type photo-MOSFETs*. They have a semitransparent gate through which the semiconductor region under the gate is illuminated. This leads to the production of electron-hole pairs in that area, thereby changing the threshold voltage at which the channel is induced, and also the transconductance which is the key parameter of this device. Sometimes, a d. c. voltage is fed to the gate so as to set the initial operating conditions.

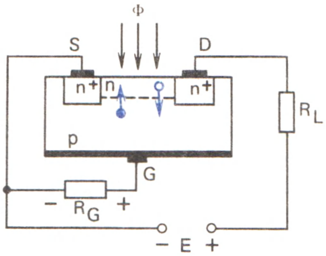


Fig. 23-12

Structure and connection of an *n*-channel photo-FET

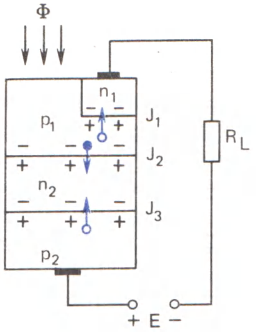


Fig. 23-13

Structure and connection of a photothyristor

Still another version of phototransistors are *photounijunction transistors* in which irradiance brings down the turn-on voltage.

23-6 Photothyristors

Four-layer *p-n-p-n* thyristor structures, such as shown in Fig. 23-13, can be controlled by light in much the same way as triode thyristors are controlled by the voltage applied to one of the emitter junctions.

The operation of a typical *photothyristor* is as follows. With the proper bias, photons incident on the p_1 base region create electron-hole pairs in that region, and these diffuse towards the *n-p* junctions. On entering the reverse-biased J_2 region, the electrons reduce its resistance. In consequence, the voltage applied to the thyristor is re-distributed: the voltage across the J_2 junction somewhat falls and the voltages across the J_1 and J_3 junctions somewhat rise. This, however, enhances carrier injection into the J_1 and J_3 junctions, the injected carriers arrive at the J_2 junction, its resistance decreases still more, and a further voltage re-distribution takes place, with a further enhancement in carrier injection in the J_1 and J_3 junctions and with the current building up cumulatively in an avalanche fashion (the dashed curves in the plot) – the thyristor turns ON.

The greater the luminous flux incident on the thyristor, the lower the voltage that is needed to turn on the device. This is clearly shown in the current-voltage characteristics of photothyristors (Fig. 23-14). After turn-on, a small voltage is established across the device as usual and nearly all of the supply voltage E is dropped across the load. Sometimes, a lead is made to one of the base regions (p_1 or n_2). By applying a d. c. voltage to the respective emitter junction via this additional lead, it is possible to bring down the turn-on voltage. The turn-on action proper will be accomplished by the incident radiation as before.

Photothyristors can successfully act as solid-state switches in automatic-control circuits handling large blocks of power and high voltages. Among the most important advantages of photothyristors are low power drain in the ON state, small size, freedom from sparking, and a short turn-on time (a split second).

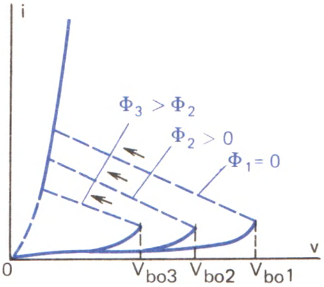


Fig. 23-14

Current-voltage characteristics of a photothyristor

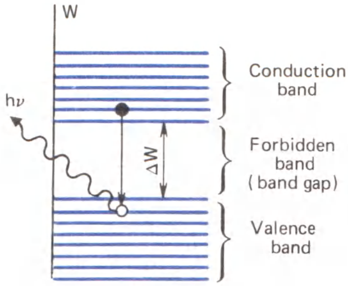


Fig. 23-15

Emission of radiation on recombination

23-7 Light-Emitting Diodes (LEDs)

Light-emitting diodes (LEDs) are electroluminescent semiconductor *p-n* junction devices that emit optical radiation when operated under forward-bias conditions. Sometimes, they are called *injection* LEDs, and the optical radiation they thus emit is classed as *injection electroluminescence.*

The earliest observations of electroluminescence were made as part of research work on radio-wave detectors. Back in 1907, Round reported that yellow light was produced when a current was passed through a silicon carbide detector. Similar results were observed by O. V. Lossev of Russian in 1923, again when working on carborundum (that is, SiC) detectors. Lossev made a systematic study of this effect and published his results in a series of papers over a period of years up to 1940. Lossev clearly recognized that the effect he had observed was connected with a rectification process. The phenomenon observed by Lossev is the basis of the work described in many books and is still often known as the *Lossev effect*. An in-depth study of the effect did not begin until the early 1950s. At present, tens of LED types are available commercially along with a wide range of more sophisticated devices made up of a varying number of LEDs.

A LED operates on the following principle. When a semiconductor diode is forward-biased, carriers are injected from the emitter region into the base region. For example, if the electron concentration in the *n*-region is higher than the hole concentration in the *p*-region, that is, if $n_n > p_p$, the injection of electrons from the *n*- into the *p*-region takes place. The injected electrons recombine with the holes of the *p*-region which are the majority carriers in the circumstances. When a free electron recombines, it may fall all the way from an unbound or higher-energy level to its ground state, releasing a photon of a wavelength corresponding to the energy-level difference associated with this transition. This ground level is situated near the top boundary of the valence band as shown in Fig. 23-15 and the energy of the released photon is nearly equal to the energy gap ΔW, that is,

$$h\upsilon = hc/\lambda \approx \Delta W \qquad (23\text{-}2)$$

On substituting the constants in Eq. (23-2), we can determine the energy gap, or the width of the forbidden band, ΔW, in electron-volts required for a photon to be emitted with any desired wavelength λ (in micrometres):

$$\Delta W \approx 1.23/\lambda \qquad (23\text{-}3)$$

It follows from the relation in (23-3) that for the emitted radiation to fall within the visible region, that is, at wavelengths from 0.38 to 0.78 μm, the semiconductor should have an energy gap of $\Delta W > 1.7$ eV. Germanium and silicon cannot be used to make LEDs because their energy gap is too narrow. State-of-the-art LEDs are mainly fabricated from gallium phosphide (GaP), silicon carbide (SiC), and some ternary compounds called *solid solutions* which consist of gallium, aluminium and arsenic (GaAlAs), or gallium, arsenic and phosphorus (GaAsP), and some others. The desired emission colour can be obtained by adding a suitable dopant to the parent semiconductor material.

Although in its narrowest sense the name 'LED' implies that the radiation emitted falls in the visible region of the spectrum, it is often convenient to broaden the term to include devices that emit in the near-infrared region of the spectrum. Infrared (or IR) light-emitting diodes are mainly fabricated from gallium arsenide, GaAs. IR LEDs are used in photorelays, various transducers, and as photodetectors in optocouplers.

There are also LEDs capable of emitting any one of two colours. They are built with two light-emitting junctions one of which has the spectral response peak in one region (say, red) of the spectrum, and the other, in another region (say, green) of the spectrum. Which colour is emitted will depend on the relative magnitude of the currents flowing through the junctions. The best performance of all is shown by LEDs incorporating heterojunctions.

The key parameters of LEDs are as follows.

1. *Radiant* (or *luminous*) *intensity*, which is measured in candelas and specified for a particular value of forward current. The radiant intensity of LEDs is usually a few tenths of a millicandela or several millicandelas. The candela is the unit of luminous intensity emitted by a special standard source.

2. *Brightness*, which is defined as the ratio between the luminous intensity and the area of the light-emitting surface. It ranges between tens and hundreds of candelas per square centimetre.

3. *Direct forward voltage* (2 or 3 V).

4. *Emission colour* and *wavelength at the peak luminous flux* (or peak emission wavelength).

The maximum ratings of LEDs include the forward continuous current (it is usually set at tens of milliamperes), the direct reverse voltage (several volts), the total device dissipation at a specified temperature, the operating and storage junction temperature range (usually from − 60° to + 70°C).

Several characteristics come into consideration in the applications involving LEDs. One is the brightness characteristic which relates brightness to forward current. The other is the luminous intensity characteristic which relates the luminous intensity of a device to its forward current. The spectral response of a LED shows its emission as a function of wavelength. The current-voltage characteristic of a LED is similar to that of a conventional rectifying diode. A very important characteristic of LEDs is their *radiation directivity pattern*; it is mainly determined by the construction of a LED, notably its lens and other features. The emission from a LED may be directional or diffuse.

Some parameters of LEDs are temperature-dependent. For example, their brightness and radiant intensity fall with rising temperature. LEDs have a high speed of response. The emission rises to its maximum in a matter of about 10^{-8} s after a forward-current pulse is applied to the device.

LEDs are configured so that as much of the output luminous flux could find its way out as practicable. Still, a sizeable proportion of the emission is lost due to the absorption in the semiconductor material itself and due to the total internal reflection at the interface between the crystal and air. Physically, LEDs are enclosed into metal cases fitted with a lens which provides for a directional radiation pattern, or in a transparent metal case which results in a diffuse emission. There are also LEDs which come uncased. The weight of a LED is a few hundredths of a gram or a little more.

LEDs are the basic building blocks of more elaborate devices. One of them is the *linear LED scale* which is essentially an IC consisting of a consecutively disposed light-emitting structures, or segments, ranging from five to 100 in number. These linear scales can replace switchboard-mounted instruments and are able to display continuously varying information.

Another LED-based device is the *alpha-numeric display* or *indicator*. It is likewise fabricated in the form of an IC containing several light-emitting structures (segments) disposed so that they can be combined into various luminous numerals or letters. A single-digit LED display can present the numerals 0 to 9 or some letters, but one at a time. Multi-digit LED displays can present several numerals, letters or symbols at a time. In most cases, the segments that make up the displayed characters are bars (there are usually seven bars or segments for each digit position). There are also *dot-matrix displays* each consisting of 35 dot-shaped LEDs. With them, any characters can be synthesized. A major advantage of the dot-matrix display using a large number of LEDs is that failure of any one of them does not result in an erroneously reproduced character. In seven-segment displays, failure of only one segment renders the device completely useless.

It is several years now that research work has been under way on multi-element blocks each containing thousands of LEDs so that intricate shapes and patterns could be displayed. Among other things, this principle is at the basis of *flat-plate display panels* for TV receivers instead of the traditional picture tubes (kinescopes).

The parameters and characteristics of alphanumeric displays are the same as are usually quoted for LEDs. The devices are widely used in instruments, automatic control, computers, digital timepieces, etc.

23-8 Optocouplers (Optoisolators)

The combination of a miniature light source and a photodetector in the same package has led to a very useful family of devices commonly referred to as *optoisolators* or *optocouplers*, although some authors prefer to call them *light source-detector combinations*. The light source converts electric signals into light signals which strike the photodetector and produce in it electric signals again. The combination of only one light source and of only one photodetector may be referred to as an *elementary optocoupler* or *optoisolator cell*. Sometimes, use is made of one or several optocouplers arranged on the same circuit board along with matching and amplifying circuits or components – the result is an *optoelectronic* (or *electro-optic*) *integrated circuit*. The signals existing at the input and output of an optocoupler (optoisolator) are always electric, while the input and output are always optical. The light source circuit is the driving (or control) section of the device, while the photodetector circuit is the controlled (or driven) section. In electronic circuits these devices are used to provide *optical coupling* and *electrical isolation*.

The key advantages that are gained from the use of optocouplers may be summed up as follows.

1. There is no electric connection between input and output and feedback between the photodetector and the light source. The isolation resistance between input and output may be as high as 10^{12}-10^{14} Ω, and the transfer capacitance does not exceed 2 pF, being as small as a fraction of a picofarad in some devices.

2. The bandwidth is very large, extending from zero to 10^{13}-10^{14} GHz.

3. The output signal can be controlled by way of the optical section of the device.

4. The optical channel is highly immune to interference that may arise due to external electromagnetic fields.

5. In an electronic equipment, optocouplers may be teamed up with other semiconductor and microelectronic devices.

On the demerit side, the following should be stressed:

1. The power drain is relatively high because energy conversion takes place twice, and the efficiency of each of them is low.

2. Optocouplers are subject to temperature variations and radiation effects.

3. There is a noticeable tendency to ageing, that is, an impairment in performance with time.

4. The inherent noise level is relatively high.

5. Optocouplers have to be manufactured by hybrid technology instead of by the more convenient and perfect planar technology – a single optocoupler combines a light source and a photodetector made of different semiconductor materials.

All of these drawbacks are gradually minimized or eliminated altogether with further progress in optoelectronics.

Physically, the light source and the photodetector of an optocoupler are enclosed in a case and bonded with an optically transparent cement (Fig. 23-16). Hybrid ICs use specially designed uncased miniature optocouplers. Standing apart from the devices where the source and detector are separated by a transparent optical couplant are optocouplers in which there is an air gap left between the source and the detector (Fig. 23-17*a*). The gap may receive a moving strip of an opaque material, such as punched tape, for control of the luminous flux. In another form of optocoupler with an exposed optical channel, the luminous flux from the light source is incident on the photodetector after being reflected from some external object (Fig. 23-17*b*).

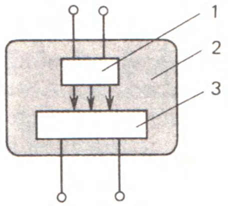


Fig. 23-16

Basic structure of an optocoupler (optoisolator): (*1*) light source; (*2*) optically transparent cement (couplant); (*3*) photodetector

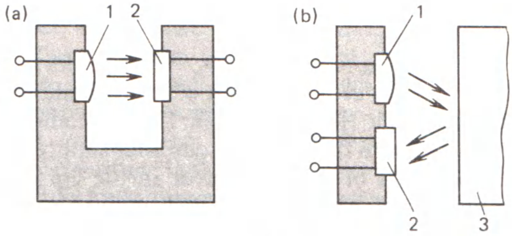


Fig. 23-17

Optocouplers with an open optical channel: (*1*) light source; (*2*) photodetector; (*3*) external object

Optocouplers are available in practically all likely combinations of light sources and photodetectors, such as a lamp and a photoresistor, a LED and a photoresistor, a LED and a photodiode, a LED and a phototransistor, a LED and a photodarlington, a LED and a photothyristor, some of these combinations coming as discrete or IC devices. Let us examine each combination briefly, referring to each by the type of photodetector used.

Photoresistor optocouplers. The light source is an extremely small incandescent lamp or a LED emitting in the visible or IR region of the spectrum. The photodetector is a cadmium-selenide (CdSe) or cadmium sulphide (CdS) photoresistor (or bulk photoconductor) for visible light emission, or lead selenide or sulphide (PbSe or PbS) photoresistor for emission in the infrared. The photoresistor is able to operate on both a. c. and d. c. For good performance of the device, it is essential to achieve a good match between the light source and the photodetector in terms of spectral response.

In sketch form the arrangement of a photoresistor optocoupler is shown in Fig. 23-18. The light source is a LED and the photodetector is a photoresistor. The output circuit operates on a direct or alternating voltage *E* and is loaded into R_L. The voltage applied to the LED controls the current in the load, and so is designated as V_{cont}. The drive section (the source circuit) is reliably isolated from the photoresistor which may therefore be connected to a circuit at a relatively high voltage, say, 220 V.

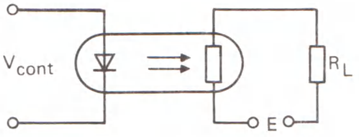


Fig. 23-18

Connection of a resistor-type optocoupler in a circuit

The parameters usually specified for this class of optocouplers are the maximum currents and voltages at the input and output, the output resistance under normal operating conditions, and what is called the *dark output resistance* (which the device presents when a dark current of several microamperes is flowing through the photoresistor in the absence of the input current), the isolation resistance, the maximum isolation voltage between input and output, the transfer capacitance, the turn-on and turn-off times which characterize the inertia of the device. The key characteristics of the devices are their input current-voltage characteristic and their current-transfer characteristic. The latter relates the output resistance to the input current.

Commercially available photoresistor optocouplers use light sources in the form of incandescent lamps, electroluminescent capacitors and LEDs. Some optocouplers intended for switching applications may use several photoresistors per device. The principal applications for photoresistor optoisolators include automa-

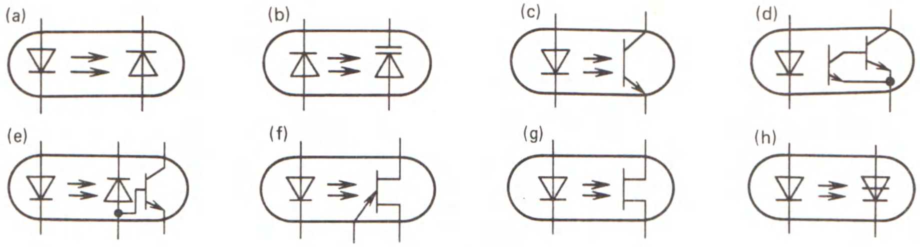


Fig. 23-19

Several combinations of light sources and photodetectors into optocouplers (optoisolators)

tic gain control, interstage coupling, and signal conditioning.

Photodiode optocouplers. These devices (Fig. 23-19*a*) usually consist of a silicon photodiode and an IR GaAs LED. The photodiode may operate in the voltage mode, thus supplying a photo-emf of 0.5-0.8 volt, or in the current mode. The diodes are fabricated by the planar-epitaxial technology. For high speed of response, the photodiodes are of the *p-i-n* type.

The key parameters of photodiode optocouplers are the input and output voltages and currents for pulsed and CW work, the current transfer ratio (CTR) defined as the ratio of the optocoupler output current to its input current in percent, the rise and fall times of the output signal, and some of the quantities quoted for photoresistor optocouplers. The CTR is usually several per cent, while the rise and fall times for *p-i-n* photodiodes may be as short as a few nanoseconds. The performance of photodiode optocouplers are described by giving their input and output current-voltage characteristics and transfer characteristics for the voltage and current modes.

Sometimes a photodiode optocoupler may have several units combined in the same case. They are then called *multichannel* optocouplers or optoisolators. The weight of a single optocoupler ranges from about one gram to several tenths of a gram. The devices are enclosed in metal-glass cases when they are intended for use as discrete components. For use in ICs, they are left uncased.

The uses to which photodiode optocouplers may be put are many and diverse. For example, they can serve as the basis for coilless pulse transformers – a feature valuable for ICs. Optocouplers are used to transfer signals between sections of electronic and telecommunication equipments, and also to control the operation of a large variety of ICs, especially MOSFET chips which use a very small input current. One more device may be looked upon as a version of the photodiode optocoupler – the one in which the photodetector is a photovaricap (Fig. 23-19*b*).

Phototransistor optocouplers. These devices (23-19*c*) usually employ a GaAs LED as the light source and a *p-n-p* bipolar silicon phototransistor as the photodetector. The leading parameters of the driving section of such devices are similar to those of photodiode optocouplers. Additionally, makers quote the maximum absolute ratings in terms of current, voltage and power for the output circuit, the dark current of the phototransistor, the turn-on and turn-off times, and the quantities describing the degree of isolation between the input and output circuits.

Phototransistor optocouplers are mainly used as switches and relays in switching circuits, couplers between transducers and indicators, to name but a few applications.

For better sensitivity, an optocoupler may use a compound-connected phototransistor (Fig. 23-19*d*) or a photodiode-transistor amplifier (Fig. 23-19*e*). Optocouplers using compound-connected phototransistors have the largest CTR but the lowest speed. On the contrary, optocouplers using a photodiode-transistor amplifier show the highest speed.

Unijunction phototransistors may likewise be employed as photodetectors in optocouplers (Fig. 23-19*f*). These devices are usually employed in switching circuits, such as those intended to control relaxation oscillators which generate rectangular pulses. A unijunction phototransistor is a versatile device: it may be used as a phototransistor if its emitter junction is left de-energized, or as a photodiode if only the emitter junction is connected in circuit.

A further version of the phototransistor optocoupler is the device using a photo-FET as the photodetector (Fig. 23-19*g*). The output current-voltage characteristic is then highly linear over a wide range of voltages and currents, and so these devices are well suited for use in analog circuits.

Photothyristor optocouplers. Devices in this class (Fig. 23-19*h*) use a silicon photothyristor (or a photo-SCR) as the photodetector and are employed solely as switches to generate strong pulses which control high-power thyristors, and also control and switch heavy loads.

The key parameters of photothyristor optocouplers are the input and output currents and voltages at turn-on, during operation, and as absolute maximum ratings, and also the turn-on and turn-off times, and the degree of isolation between input and output.

Integrated optocouplers. These are optical coupling between the individual elements and electrical isolation between them. ICs using photodiodes, phototransistors or photothyristors consist of light sources, photodetectors, and circuits to handle the signals generated by the photodetectors. A distinction of optoelectronic ICs is unilateral signal transfer and complete elimination of feedback.

Optoelectronic ICs are used to handle logic and analog signals, to operate as solid-state relays, and in alpha-numeric display circuits. In addition to the usual parameters associated with discrete optocouplers, their IC counterparts are characterized by stating their input and output currents and voltages corresponding to logic 0 and 1, turn-on and turn-off delay times, supply voltage and current.

There are also other types of optocouplers, both discrete and integrated, which have, for example, an optical input and an optical output and are intended to handle optical signals, optoelectronic indicator ICs with several built-in LEDs or segment-type character display elements. Optoelectronics is a very promising field which is expanding all the time.

Conclusion

The electron devices, both semiconductor and tube-type, we have examined in this book do not cover all the likely designs, types, and makes. Still, they give a fairly accurate picture of the state of the art. With knowledge of their construction, operation, characteristics, parameters, basic properties and uses, the reader will be able to cope with the construction and operation of other similar devices on his own.

Electronics is advancing at a breath-taking pace. Ever new devices are appearing almost every day for ever higher frequencies (far into the gigahertz range), for ever greater powers, and for ever higher temperatures while their size is made progressively smaller. A good deal of emphasis is placed on improvements in reliability, durability, stability, ruggedness, temperature and radiation resistance. The advance is especially rapid in microelectronics and quantum electronics, and many more specialized divisions are added to the field.

One such division is *acoustoelectronics* which utilizes the interaction of ultrasonic waves with electrons in metals and semiconductors. Acoustoelectronic devices are successfully used to process radio signals. Examples are *acoustic wave devices* which depend for their operation on so-called surface acoustic waves propagated along the surface of a substrate. The associated electric field extends for a short distance out of the surface and can interact with the conduction electrons of a separate semiconductor placed just above the surface.

Closely associated with acoustoelectronics is *piezoelectronics* known for a fairly long time now. For their operation piezoelectronic devices depend on the *piezoelectric effect* in solids – an effect that occurs when certain materials are subjected to mechanical stress. An electrical polarization is set up in the crystal, and the faces of the crystal become electrically charged. The polarity of the charges reverses if the compression is changed to tension. There is also the *reverse piezoelectric effect* which occurs when an electric field is applied across the material and causes it to contract or expand according to the sign of the applied electric field.

A further division of electronics has to do with *solions* – electrochemical sensing and control devices in which ions in solution carry electric charges to give amplification corresponding to that of vacuum tubes and transistors. A solion consists of two or more electrodes sealed in an electrolyte. Solions are low-frequency devices because of the appreciable inertia of ions. Yet, where they are applicable, they offer a reduction in power drain, simplify the associated circuitry, enhance reliability and ruggedness. In some cases, a single solion is able to do the job of a complete circuit assembly. The basic solion devices are the diode, the integrator, the pressure transducer, and the electroosmotic driver. More sophisticated designs of solions can perform various mathematical, process-control, and data-processing functions. What is especially valuable about solions is that they can readily be re-adapted to perform quite different duties.

An ever wider field is being gained by *cryotronics* (which is a contraction of 'cryogenic electronics'). It has to do with devices operating at extremely low (*cryogenic*) temperatures. The key device is the *cryotron*, a superconductive two-port (four-terminal) device in which a magnetic field, produced by the flow of a current through the input port, controls the superconducting-to-normal transition and, in consequence, the resistance at the output port (that is, between the two output terminals).

These and other divisions of state-of-the-art electronics owe their existence to veritable breakthroughs in solid-state and quantum physics. Undoubtedly, many more exciting discoveries will be made in the field of electronics, giving further impetus to progress in science and technology in general.

Index